Université de Montréal

Parameter-Efficient Modeling and Robust Automatic Evaluation of Image Captioning

par

Saba Ahmadi

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Informatique

October 31, 2023

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Parameter-Efficient Modeling and Robust Automatic Evaluation of Image Captioning

présenté par

Saba Ahmadi

a été évalué par un jury composé des personnes suivantes :

Irina Rish

(président-rapporteur)

Aishwarya Agrawal

(directeur de recherche)

Christopher Pal

(membre du jury)

Résumé

Le sous-titrage d'images est la tâche de l'intelligence artificielle (IA) qui consiste à décrire des images en langage naturel. Cette tâche d'IA a plusieurs applications sociétales utiles, telles que l'accessibilité pour les malvoyants, la génération automatisée de contenu, l'interaction humain-robot et l'analyse d'imagerie médicale. Au cours des huit dernières années, la recherche sur le sous-titrage d'images a connu d'énormes progrès dans la création de modèles solides, la collecte d'ensembles de données à grande échelle ainsi que le développement de mesures d'évaluation automatique.

Malgré ces progrès remarquables, la recherche sur le sous-titrage d'images est confrontée à deux défis majeurs: 1) Comment construire des modèles efficaces en termes de paramètres, et 2) Comment construire des métriques d'évaluation automatique robustes. Dans cette thèse, nous apportons notre contribution à la résolution de chacun de ces défis.

Premièrement, nous proposons une méthode efficace en termes de paramètres (MAPL [**65**]) qui adapte des modèles pré-entraînés unimodaux de vision uniquement et de langage uniquement pour la tâche multimodale de sous-titrage d'images. MAPL apprend un mappage léger entre les espaces de représentation des modèles unimodaux. Ainsi, MAPL peut exploiter les fortes capacités de généralisation des modèles unimodaux pré-entraînés pour des tâches multimodales telles que le sous-titrage d'images.

Deuxièmement, nous présentons une étude systématique de la robustesse des mesures d'évaluation des sous-titres d'images récemment proposées. Même si ces métriques correspondent bien aux jugements humains, nous avons constaté qu'elles ne sont pas robustes pour identifier les erreurs fines dans les légendes générées par le modèle. Il faut donc faire preuve de prudence lors de l'utilisation de ces métriques pour l'évaluation des sous-titres d'images. Nous espérons que nos résultats guideront de nouvelles améliorations dans l'évaluation automatique du sous-titrage d'images.

Mots-clés : paramètre-efficace, sous-titrage d'images, évaluation, métriques, sans référence

Abstract

Image captioning is the artificial intelligence (AI) task of describing images in natural language. This AI task has several useful societal applications, such as accessibility for the visually impaired, automated content generation, human-robot interaction, and medical imaging analysis. Over the last eight years, image captioning research has seen tremendous progress in building strong models, collecting large scale datasets as well as developing automatic evaluation metrics.

Despite such remarkable progress, image captioning research faces two major challenges: 1) How to build parameter-efficient models, and 2) How to build robust automatic evaluation metrics. In this thesis, we make contributions towards tackling each of these challenges.

First, we propose a parameter efficient method (MAPL [**65**]) that adapts pre-trained unimodal vision-only and language-only models for the multimodal task of image captioning. MAPL learns a lightweight mapping between the representation spaces of the unimodal models. Thus, MAPL can leverage the strong generalization capabilities of the pre-trained unimodal models for multimodal tasks such as image captioning.

Second, we present a systematic study of the robustness of recently proposed image captioning evaluation metrics. Even though these metrics correlate well with human judgments, we found that these metrics are not robust in identifying fine-grained errors in model generated captions, and thus, caution needs to be exercised when using these metrics for image captioning evaluation. We hope our findings will guide further improvements in the automatic evaluation of image captioning.

Keywords: parameter-efficient, image captioning, evaluation, metrics, reference-free

Contents

List of tables

List of figures

List of acronyms and abbreviations

AI	Artificial Intelligence
NN	Neural Networks
VQA	Visual Question Answering
CNN	Convolutional Neural Network
FC	Fully Connected Layer
IID	Independent and Identically Distributed
OOD	Out-of-Distribution
LSTM	Long Short-Term Memory
VL	Vision-Language
VLMs	Vision-Language Models
AGI	Artificial General Intelligence

LLMs	Large Language Models

Acknowledgement

I extend my heartfelt gratitude to those who generously supported me during my master's journey. I dedicate this thesis to my dear mother, who sadly left us from Covid-19 at the start of my research journey. Her constant love and support have been my guiding light, and this work is a tribute to the lasting impact she had on my academic pursuits.

Foremost, I express my deepest appreciation to my supervisor, Aishwarya Agrawal, who not only played a pivotal role in shaping my academic journey but also unwaveringly stood by me throughout this transformative experience. I express my sincere gratitude to Aishwarya for accepting me as her student. Under her guidance, I acquired invaluable skills in defining research problems, honed my research methodology, and developed critical thinking abilities. Aishwarya's commitment to my academic growth was evident through her meticulous feedback and patient explanations, ensuring my comprehensive understanding. Even amidst her demanding schedule, including travels for conferences with time zone differences, she dedicated late hours to provide detailed feedback on my paper writing, aiding me in improving my articulation and explanatory skills. In moments of error, she approaches challenges as a team, diligently aiding in problem resolution and fostering an environment where learning from mistakes is encouraged. I am profoundly thankful for Aishwarya's tireless support, encouragement to pursue my passions, and dedication to fostering my learning journey.

Moreover, Aishwarya's considerate treatment of her students has cultivated a sense of belonging akin to a familial bond. Even in matters unrelated to research, she extends her support, creating a nurturing space for addressing various challenges. Throughout my master's program, I faced numerous personal and academic obstacles that seemed insurmountable. Aishwarya's steadfast encouragement during the emotionally and mentally challenging period following the loss of my mother due to COVID-19, remains etched in my memory. In addition to guiding me through the complexities of research, critical thinking, and writing, she imparted life lessons on resilience, compassion, and community engagement. Aishwarya's mentorship extended beyond academia, teaching me not just how to be a better researcher but also a better person who listens and contributes positively to the community.

I am deeply grateful for the invaluable assistance I received when starting my graduate school journey, particularly from my labmate, Oscar Mañas. Our collaboration on my initial research project was pivotal, and I extend my sincere appreciation to Oscar for graciously allowing me to join his research and for patiently explaining various aspects to me. Additionally, I extend my gratitude to our collaborators Pau Rodriguez Lopez, Yash Goyal, and Aida Nematzadeh — all of whom contributed significantly to my learning through their constructive feedback.

I express my gratitude to my labmates, Rabiul Awal, Le Zhang, and Sarvjeet Singh Ghotra, for their insightful discussions and constructive feedback. I extend my genuine appreciation for the positive and collaborative atmosphere they have fostered within our lab, which has significantly contributed to my growth and learning experience.

I extend my gratitude to the entire Mila community for fostering an enriching environment and providing robust infrastructure. The technical support from the Mila IDT team in managing the computational infrastructure is sincerely appreciated. I would also like to express my thanks to Olexa Bilaniuk for his dedicated efforts and timely responses, not just for resolving technical issues but also for taking the time to explain them in a way that allows me to understand and learn.

I sincerely want to express my gratitude to my friends; this master's journey would not have been possible without your support. A special thanks to (sorted alphabetically) Milad Aghajohari, Kamran Chitsaz, Beheshteh Halimi, Amirhosein Kazemnezhad, Amin Memarian, Rozhin Nobahari, and Maryam Saeidi. Your support has meant the world to me, and I am genuinely appreciative.

I express my heartfelt gratitude to my family, beginning with my mother, Sona Adham, who made immense sacrifices to help me pursue my dreams. She not only served as my mother but also as a role model, consistently encouraging me to chase my aspirations while being realistic and finding joy in the journey, irrespective of the final outcome. My sister, Sara Ahmadi, has been a constant source of love and support, serving as another inspiring figure in my life through all the ups and downs. Thanks to my father, Ali Ahmadi, whose kindness and belief in me have been a pillar of support. A special acknowledgment goes to my dear partner, Amin Darabi. Your dedicated support over the years has been invaluable, and I am grateful to have such an intelligent partner with whom I can discuss my work and ideas.

I also extend my thanks to the rest of my family, friends and colleagues who I have not mentioned here. I feel truly fortunate to be surrounded by such a supportive and kind group of people like you.

Introduction

One objective of Artificial Intelligence (AI) is to create systems capable of visually perceiving images (comprehending the content of an image, including identifying individuals, activities, and locations) and expressing this understanding to humans using natural language. These systems are referred to as vision-language (VL) systems and demand the modeling of multimodal data. These systems must be adept at joint modeling of visual and textual data. Applications of such systems span various scenarios:

- Aiding individuals with visual impairments (User: "Describe the painting in front of me," AI: "It depicts a lively scene with rolling hills under a clear blue sky."),
- Facilitating engaging learning experiences for children (AI: "This is a simulation of the solar system. The bright object in the center is the sun and the smaller objects orbiting it are planets."),
- Simplifying online shopping through natural language queries (User: "Find me a floral mini-dress with a blue background."),
- Elevating interactions with personal home robots (User: "Did you notice when our red car exited the garage?").

In recent years, there has been significant interest in vision and language models driven by their pivotal applications. Beyond their evident societal applications, the acquisition of multimodal vision-language (VL) learning stands as an indispensable milestone on the path toward artificial general intelligence (AGI).

Although significant progress has been made in vision and language research, numerous challenges persist, motivating ongoing efforts for improvement. This thesis focuses on addressing challenges in the vision and language field, particularly in the domain of image captioning. Our primary objectives include overcoming the demand for parameter-efficient learning, which entails developing models that can learn with fewer parameters, enabling them to operate in resource-constrained environments. Additionally, we target the formidable challenge of robust automatic evaluation, as it involves developing evaluation metrics that can accurately assess the performance of vision and language models, especially when human annotation is unavailable.

In a broader picture, our work intersects with challenges such as out-of-distribution generalization, which involves training vision and language models to generalize beyond the data distribution seen during training. We also delve into data-efficient learning that involves adapting a pre-existing model to a novel task or a new dataset within the same task using a limited number of samples. Another facet of our research involves tackling the challenges of interpretability and explainability, which involves developing models that can provide interpretable and explainable outputs, enabling users to understand how the model arrived at its output.

Below is an overview of the specific dimensions of image captioning that I study in this thesis.

0.1. MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting (Chapter 2)

In our first work, MAPL [**65**], we develop a parameter-efficient method, MAPL, for repurposing pre-trained unimodal models for multimodal tasks. In this method, we learn a lightweight mapping between the representation spaces of the pre-trained vision encoder and the pre-trained language model using aligned image-text data. By doing so, it can generalize to unseen vision-language tasks with only a few in-context examples, making it effective for low-data and in-domain learning. In addition, our model reuses frozen vision-only and language-only foundation models; hence, the expensive computational resources used to train these models can be saved to help reduce energy and carbon costs.

0.2. An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics (Chapter 3)

In our second work, we examine CLIPScore [**37**], UMIC [**50**], and PAC-S [**79**], three recently proposed image captioning metrics. We carried out controlled experiments to assess the responsiveness of these metrics to diverse visual and linguistic aspects, including fine-grained distinctions between captions, the count and size of objects in images referenced in captions, sentence structure, and negation of captions. The goal of this study is to assess the robustness and sensitivity of these metrics in various contexts, as well as to pinpoint opportunities for enhancing reference-free evaluation in the field of image captioning.

0.3. Our Contribution

This thesis addresses and explores primary challenges in the field of image captioning, which are outlined as follows:

- In our initial work, MAPL, we addressed specific challenges within the image captioning domain:
    - Parameter-Efficient Learning Challenge: Strategic reuse of frozen vision-only and language-only foundation models in our approach, which leads to Parameter-Efficient Learning.
    - Data-Efficient Learning Challenge: Achieved generalization to novel vision-language tasks with only a few in-context examples.
    - Out-of-Distribution Generalization Challenge: Developed a lightweight mapping between pre-trained vision encoders and language models using aligned image-text data, which demonstrated effective generalization to new vision-language tasks with minimal in-context examples and also generalization to the same task but for other datasets.
- In our second study, where we examine the robustness of three recently introduced metrics for image captioning, we made contributions to two critical challenges:
    - Robust Automatic Evaluation Challenge: Conducted a thorough assessment of the robustness and sensitivity of three recently proposed image captioning metrics across diverse visual and linguistic contexts.
    - Interpretability and Explainability Challenge: Delved into an examination of three recently proposed image captioning metrics, evaluating their responsiveness to various visual and linguistic aspects (such as fine-grained distinctions between captions, the count and size of objects referenced in images, sentence structure, and the presence of negation in captions) and provided insights that empower users of these metrics to comprehend how these factors influence the assigned scores by these metrics, guiding them on when to exercise caution in employing these metrics.

0.4. Thesis layout

In the first chapter, we supply the necessary background knowledge for comprehending the two works presented in this thesis. We delve into the architectures of vision and language models, covering aspects such as pre-training, transfer to downstream tasks, and image captioning evaluation. In the second chapter, we introduce our work titled 'MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting'. The third chapter presents our work, 'An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics'. Lastly, chapter four includes conclusions derived from this thesis, along with suggestions for future research endeavors.

Chapter 1

Background

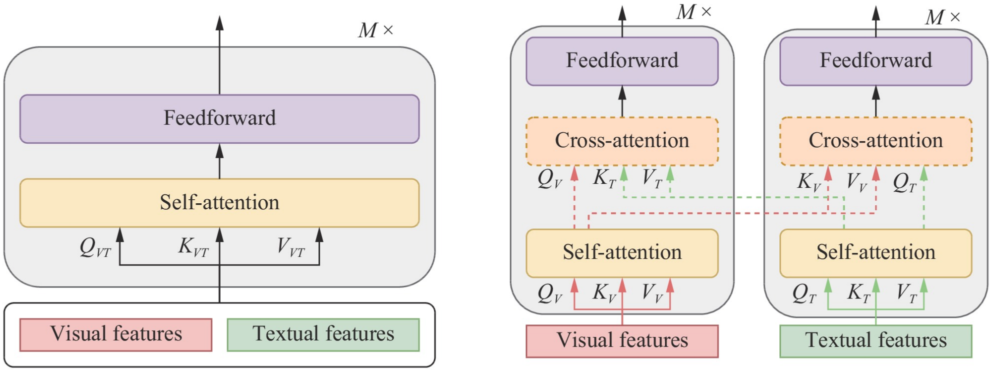
In this chapter, we delve into the detailed technical background required to comprehend the architecture of vision-language models, and vision-language models' training and evaluation.

1.1. Vision-language models architecture

Building on the remarkable success of transformers in natural language processing, researchers have extended this approach to multimodal contexts, such as tasks involving vision and language. Architectures for vision and language models generally fall into two categories. The first category encompasses models that independently process each modality of vision and language, followed by a distinct module that integrates these modalities. In contrast, the second category comprises models that simultaneously accept image and language inputs, learning a unified representation for both concurrently. Subsequently, we will provide a detailed explanation of each architecture and its corresponding components.

- **Dual-encoder**: Dual encoder models such as ALIGN[**42**] and CLIP [**71**]) comprise two essential components: a vision-encoder and a text-encoder. Each encoder is specifically designed to learn representations of vision-only and text-only modalities independently. The two learned representations are then projected into the same space and then, using a simple operation such as dot product, can be interpreted as a similarity score for two modalities (exemplified by CLIP [**71**]). These models excel in image-retrieval tasks and, when trained with extensive data and techniques like contrastive learning, exhibit exceptional performance as vision models. However, since the two modalities are learned separately, these models tend to underperform in tasks that necessitate a more complicated interpretation of vision and language together, such as visual question answering.

- **Unified model**: To handle complex tasks that necessitate deep interactions between two modalities, models often incorporate a vision-encoder, text-encoder, and additional transformer layers to learn the interplay between the two modalities. There are two different ways to use transformers for vision and language tasks:
  (1) Single-stream: Models such as VisualBERT [**54**], VLP [**112**], VL-BERT [**87**], OSCAR [**55**], UNITER [**20**], VinVL [**108**] and ViLT [**46**] utilize a single stack of transformers to model both the vision and text modalities. These models concatenate the text and visual features and feed them into a single transformer block, allowing the transformer to learn relationships between the two modalities.
  (2) Dual-streams: Models such as ViLBERT [**60**], LXMERT [**90**], and METER [**28**] employ a two-stream architecture in which text and visual features are input into separate transformer blocks independently. Techniques such as cross-attention, alongside self-attention, are utilized to learn the interactions between two modalities.



(a) Single-stream architecture

(b) Dual-stream architecture

Fig. 1.1. This figure is from [**15**] and demonstrates the architecture of single stream and two streams unified vision and language models.

In the following, our primary objective is to describe the essential components of vision and language models: the vision-encoder, the text-encoder and large language models.

1.1.1. Vision-encoder

There are primarily three types of vision encoders employed to extract image features:

- **CNNs:** Convolutional Neural Networks (CNNs) are commonly used to extract features from grid-like data. Typical CNN architecture includes convolution and pooling layers, followed by one or more fully connected layers. Some of the widely used CNN architectures for feature extraction include AlexNet [**48**], VGGNet [**85**], GoogLeNet [**89**], and ResNet [**36**].

There are two primary ways to utilize the learned representation of an image from a CNN. One approach involves using **grid features** from earlier layers, which allows for incorporating spatial and local information in the image. In contrast, the other approach employs **global features** obtained after fully connected layers, which learn visual features of the entire image. This method may result in capturing more salient features of the image as a whole at the expense of losing local information about features of the image. Recent models, including PixelBERT [**40**], SOHO [**39**], and SimVLM [**101**], mainly employ various ResNet [**36**] variants as their vision-encoder. CLIP-ViL [**83**] utilizes versions of CLIP [**71**] with different ResNet backbones. Utilizing CLIP as a vision encoder offers advantages over traditional CNNs. CLIP comprises a visual encoder and a text encoder that independently encodes input images and text. The dot product between the two embeddings is then employed as the similarity score between the input image and text. CLIP is pre-trained using a contrastive loss, where the model aims to maximize similarity for positive pairs (where image and caption match) and minimize it for randomly selected negative samples. pre-trained on 400 million image-text pairs from the internet, CLIP has a significantly richer vocabulary compared to the common method of training ResNet models on image classification datasets like ImageNet [**78**], which only encompass a limited range of visual concepts.

- **Object Detectors:** Models like VinVL [**108**], VisualBERT [**54**], ViLBERT [**60**], LXMERT [**90**], and UNITER [**20**] depend on object detectors to extract information from images, making the quality of the underlying object detector a critical factor in their performance, also extracting region features can also be a time-consuming process. Object detectors recognize objects in an image and output labels along with corresponding bounding boxes. One of the most widely used object detectors is the Faster R-CNN [**75**], which is pre-trained on the Visual Genome dataset [**47**]. Utilizing object detectors allows for extracting region features at the object level, which can be highly beneficial for complex vision and language tasks such as VQA.
- **Vision Transformers:** Various vision transformers have been employed in vision and language tasks, including plain ViT [**27**], DeiT [**93**], BEiT [**9**], CLIP-ViT [**71**], Swin Transformer [**59**], DINO [**14**], and MAE [**35**]. The first Vision Transformer (ViT) initially divides an image into patches, which are then flattened and linearly projected to a lower-dimension. Positional embeddings are subsequently added to the beginning of each of these embeddings. To facilitate learning the image representation, a learnable special token [CLS] is appended to the start of these embeddings. Theses embeddings are then fed into a standard transformer [**97**]. Vision transformers are pre-trained on extensive datasets, enabling them to acquire robust visual

representations. In our work, MAPL [**65**], we utilize grid features from CLIP [**71**] with ViT [**27**] backbone.

1.1.2. Text-encoder

Various types of text modules are used in vision and language models. The first type, **encoder-only**, is inspired by BERT [**25**]. First, words are divided into subwords and tokenized, and special tokens are added at the beginning and end of the tokens sequence. These tokens then pass through transformer layers based on the model's architecture, whether a dual-encoder or unified model. Models that perform generation tasks, such as captioning, use encoder-only modules to generate text sequences token by token with a causal mask. Encoder-only models are pre-trained with a masked language modeling objective that uses bi-directional attention to predict missing words, and therefore, generation tasks utilize causal masks. Models like BLIP [**51**] employ an **encoder-decoder** architecture for generation, where the decoder generates text autoregressively using both previously generated tokens and the encoder's generated representations.

1.1.3. Large Language Models

Recent models such as Flamingo [**4**] and MAPL [**65**] use large language models (LLMs). LLMs are models pre-trained on large-scale unsupervised text data and finetuned for specific tasks such as text classification, sentiment analysis, question-answering, and summarization. In Flamingo [**4**] and MAPL [**65**] the LLM can act as both encoder and decoder, generating text by attending to previous tokens only and predicting the next token.

1.2. Vision-language training and evaluation

1.2.1. Pre-training techniques

Training a model for a particular task acquires a specialized representation tailored to that specific task's requirements. However, despite the diversity of tasks, there often exists a significant overlap in the fundamental understanding of vision and language. pre-training approaches capitalize on this shared foundation by training models using objectives designed to cultivate a versatile representation of both images and language. This pre-trained model, typically trained on extensive datasets, serves as a knowledge-rich starting point. Subsequently, usually fine-tuning is applied on task-specific datasets to adapt the model to the particular downstream task. This approach yields substantial energy savings since the model already possesses a strong foundation in general vision and language comprehension, facilitating the fine-tuning process and enhancing overall efficiency. In the pre-training phase, models typically train with multiple objectives. Here, we introduce some of the most commonly used objectives for Vision-Language Models (VLMs).

- **Contrastive learning:** Contrastive learning for pre-trained vision and language models was first used by the CLIP [**71**] model and later used by models such as ALIGN [**42**] and ALBEF [**53**]. Contrastive learning aims to have image and text representations in the same space, aiming to minimize the distance between matched pairs and maximize the distance between negative samples, which are essentially randomly unmatched image-text pairs. Different models define distance in various ways. For example, CLIP projects the learned image and text representations into the same space and calculates the cosine similarity between them. Thus, for a batch size of N, the objective is to maximize the similarity between the N matched image-text pairs and minimize the similarity between the $N^2 - N$ unmatched pairs. One advantage of CLIP is that it is trained on a vast dataset of 400 million image-text pairs gathered from the internet. By being supervised with rich natural language, it can learn a substantial number of visual concepts.
- **Image-text matching (ITM):** Image-text matching is often viewed as a classification task, where the model must predict whether image and text pairs are matched (positive samples) or unmatched (negative samples). A special [CLS] token is appended to the beginning of the embeddings to learn a joint representation for the image-text pair, which is then passed through a linear layer for binary classification to determine if the image-text pairs are matched or not. This loss is crucial for training an effective visually grounded text-encoder.
- **Masked language modeling (MLM):** In this approach, the model is trained with the MLM objective, where it is presented with captions containing randomly masked words, and its task is to predict these masked word tokens. This objective encourages the model to generate image descriptions grounded on visual context from the image.

1.2.2. Transfer to downstream tasks

Transferring knowledge from pre-trained models to downstream tasks involves several strategies to make the most of the prelearned information. The choice of strategy depends on the specific characteristics of the pre-trained model, the nature of the downstream task, the availability of data, and computational resources. Effective knowledge transfer enables models to leverage existing knowledge and adapt to new challenges efficiently. In this section we discuss two main approaches, fine-tuning and few-shot prompting.

- **Fine-tuning**: The objective of fine-tuning is to adapt the pre-trained model's knowledge and representations to perform well on the new, task-specific data without training a model from scratch.

- **In context few-shot prompting:** As demonstrated by PICa [**105**], Frozen [**95**], MAPL [**65**], and Flamingo [**4**], Vision-Language Models (VLMs) have the capability to rapidly adapt to new downstream tasks through a technique known as few-shot prompting. This approach allows the model to acquire proficiency even with minimal in-context examples. For instance, consider a scenario where a model originally pre-trained for captioning is presented with only a handful of Visual Question Answering (VQA) support examples. Remarkably, through these few in-context examples, the model can learn that it should provide answers to questions related to the presented image, showcasing the remarkable adaptability and versatility of few-shot prompting in transfer learning scenarios.

  Nonetheless, it is important to note that while few-shot prompting offers adaptability, its performance may not match that of direct fine-tuning on the task. However, this approach remains highly efficient on multiple fronts. Firstly, it doesn't necessitate parameter updating, which can be resource-intensive. Secondly, few-shot prompting shines in scenarios where data for a specific task is scarce, offering a pragmatic solution to training models effectively in data-scarce environments. It strikes a valuable balance between performance and efficiency, making it a compelling option in various practical applications.

1.2.3. Evaluation

While evaluation may appear straightforward at first glance, it becomes notably challenging for tasks that are open-ended and have multiple valid answers. Take, for instance, the captioning task, where numerous sentences can effectively describe an image, even without any common words.

In this section, we provide a brief overview of how the community typically approaches the evaluation of image captioning task. However, it is essential to recognize that these evaluation methods come with their own limitations, which we will delve into more comprehensively in Chapter 3, particularly concerning the evaluation of image captioning.

Traditionally, image caption quality has been automatically evaluated through a reference-based approach, which compares generated captions to a set of reference captions provided by human annotators. The majority of automatic evaluation metrics for captioning, such as BLEU [**68**], ROUGE [**57**], CIDEr [**98**], and METEOR [**8**], compute n-gram (set of n contiguous words) matches between candidate and reference captions (measuring lexical overlap).

However, this approach can be limiting, as it does not necessarily capture the full range of acceptable captions for a given image. Moreover, it suffers from the issue of high scores being awarded to captions that employ similar vocabularies but possess vastly different semantic meanings. To address these limitations, recent studies like CLIPScore [**37**], UMIC [**50**], and PAC-S [**79**] have proposed reference-free approaches for evaluating caption quality, which more closely align with how humans judge captions. These approaches leverage large pre-trained image-text matching models to generate a score that measures the similarity between the provided image and the candidate caption.

Chapter 2

MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting

2.1. Prologue to Paper

2.1.1. Paper Details

MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting, Oscar Mañas, Pau Rodríguez*, Saba Ahmadi*, Aida Nematzadeh, Yash Goyal, Aishwarya Agrawal.

- Please note that '*' denotes equal contribution.

This paper [**65**] is published at the European Chapter of the Association for Computational Linguistics (EACL), 2023.

2.1.2. My Contributions

My contributions included implementing some baselines, setting up training and evaluation pipelines for some settings, analyzing the model outputs by creating visualizations of model output distributions, and helping with paper writing.

2.2. Abstract

Large pre-trained models have proved to be remarkable zero- and (prompt-based) few-shot learners in unimodal vision and language tasks. We propose MAPL, a simple and parameter-efficient method that reuses frozen pre-trained unimodal models and leverages their strong generalization capabilities in multimodal vision-language (VL) settings. MAPL learns a lightweight mapping between the representation spaces of unimodal models using aligned image-text data, and can generalize to unseen VL tasks from just a few in-context examples. The small number of trainable parameters makes MAPL effective at low-data and in-domain learning. Moreover, MAPL's modularity enables easy extension to other pre-trained models. Extensive experiments on several visual question answering and image captioning benchmarks show that MAPL achieves superior or competitive performance compared to similar methods while training orders of magnitude fewer parameters. MAPL can be trained in just a few hours using modest computational resources and public datasets. We release our code and pre-trained model weights at `https://github.com/mair-lab/mapl`.

2.3. Introduction

Over the past few years, natural language processing and computer vision have witnessed impressive progress in learning models capable of transferring to unseen tasks or benchmarks [**11, 110, 70, 41**]. Recently referred to as foundation models [**10**], these can be adapted to a wide range of *unimodal* vision and language tasks without any additional training.

In this work, we study reusing such powerful *unimodal* foundation models for *multimodal* vision-language (VL) downstream tasks. In particular, we propose to connect a vision encoder, such as CLIP [**71**], to an autoregressive language model (LM), such as GPT [**72, 73, 11**], with minimal additional training on multimodal data. Our goal is to obtain a single VL model that can leverage the in-context learning abilities [**11**] of the pre-trained LM to generalize to unseen VL tasks from just a few examples.

One challenge in connecting vision encoders with LMs is aligning the visual and textual representation spaces. Recent works have approached this by adapting the LM to visual representations, either by fine-tuning the entire LM [**23**] or training adapter layers [**29, 5**]. These systems are computationally expensive to train as they have a large number of learnable parameters (hundreds of millions to a few billions) and use large-scale multimodal training data. On the other hand, Frozen [**94**] keeps the LM frozen, thus learning $\sim$10× less parameters than the above methods. However, it requires training a visual encoder from scratch, which is also computationally expensive.

Differently, we aim to reuse large pre-trained unimodal models while keeping them completely frozen and free of adapter layers. We present **MAPL** (**M**ultimodal **A**daptation of **P**re-trained vision and **L**anguage models), a simple and parameter-efficient VL model capable of tackling unseen VL tasks. MAPL learns a lightweight mapping between the representation spaces of pre-trained unimodal models. MAPL has orders of magnitude fewer parameters than previous methods (including Frozen) and can be trained in just a few hours. Moreover, MAPL's modularity makes it general-purpose and easily extensible to newer and/or better pre-trained models. We evaluate MAPL on various image captioning and visual question answering (VQA) benchmarks and compare with Frozen [**94**] in a controlled setup. MAPL significantly outperforms Frozen and achieves competitive performance compared to other methods [**29, 23**] trained on comparably sized data.

We further investigate the parameter efficiency of MAPL by training on only 1% of multimodal data (thousands of examples); we call this setting *low-data* learning. We also study *in-domain* learning: training on image-text pairs from the same domain as the downstream task domains. We train MAPL directly on 100% and 1% of in-domain data for each downstream task, without first pre-training on large-scale domain-agnostic data. Thus, we train specialized versions of MAPL for each downstream domain. Such low-data and in-domain learning are particularly useful when it is difficult to pre-train on large-scale domain-agnostic data. We found MAPL to be more effective than Frozen trained under the same settings.

To summarize, our contributions are: 1) we introduce MAPL, a parameter-efficient method capable of tackling unseen VL tasks, which can be trained using only modest computational resources and public datasets; 2) we conduct extensive experiments spanning various image captioning and VQA benchmarks, demonstrating MAPL achieves superior or competitive performance compared to similar methods while training orders of magnitude fewer parameters; and 3) we further investigate the parameter-efficiency of MAPL in two settings: low-data and in-domain. Our experiments show that MAPL is more effective than the considered methods in both settings.

2.4. Related Work

Fine-tuning based VL methods. A popular family of VL methods are based on the pre-training + fine-tuning paradigm. These methods are either encoder-only [**61, 91, 18, 56, 109**] or encoder-decoder methods [**21, 102, 43, 52**] and use transformer-based architectures. These transformers are first pre-trained on domain-agnostic image-text pairs (e.g., Conceptual Captions [**81**]) using self-supervised objectives, and then fine-tuned for each downstream task (e.g., VQA, image captioning). More recent models that are designed specifically for the task of image captioning use large pre-trained LMs (e.g., GPT-2 [**73**]) and fine-tune these models with image-caption pairs [**16, 67, 62**]. While all these approaches yield state-of-the-art performance for the tasks they are fine-tuned on, the learned model weights are highly specialized for a single task and cannot transfer to new tasks with zero or few examples. Differently, MAPL reuses the same set of weights for all downstream tasks without any additional training.

Few-shot learning based VL methods. Most similar to MAPL are methods that tackle unseen VL tasks in a zero/few-shot manner, by leveraging the in-context learning abilities of large pre-trained LMs (e.g., GPT-3 [**11**]. These methods connect a vision encoder with a pre-trained LM to tackle VL tasks. Some methods [**23, 34**] achieve this connection by fine-tuning the entire LM on image-text data, while others only train adapter layers inserted into the LM [**29**]. The vision encoder is pre-trained and kept frozen in both cases. Concurrent work Flamingo [**5**] pushes this idea even further by scaling up the amount of training data and the LM size. While inserting adapter layers requires training fewer parameters compared to fine-tuning the entire LM, the number of trainable parameters is still >100M; in contrast, MAPL only has 3.4M trainable parameters. Additionally, inserting adapter layers is not straightforward since it requires modifying the computational graph of the LM; MAPL only adds an external mapping network, which is easier to incorporate on top of pre-trained models. On the other hand, Frozen [**94**] keeps the pre-trained LM frozen and instead trains a vision encoder from scratch. This approach does not scale well with larger vision encoders (Sec. 2.6.5). MAPL keeps both the vision encoder and the LM frozen (thus further reducing the number of trainable parameters) and only learns a lightweight mapping network to connect both frozen models. Similar to MAPL, concurrent work LiMBeR [**66**] also proposes to connect a frozen vision encoder with a frozen LM but using a linear mapping, which is not as parameter- and compute-efficient as MAPL (Sec. 2.6.5).

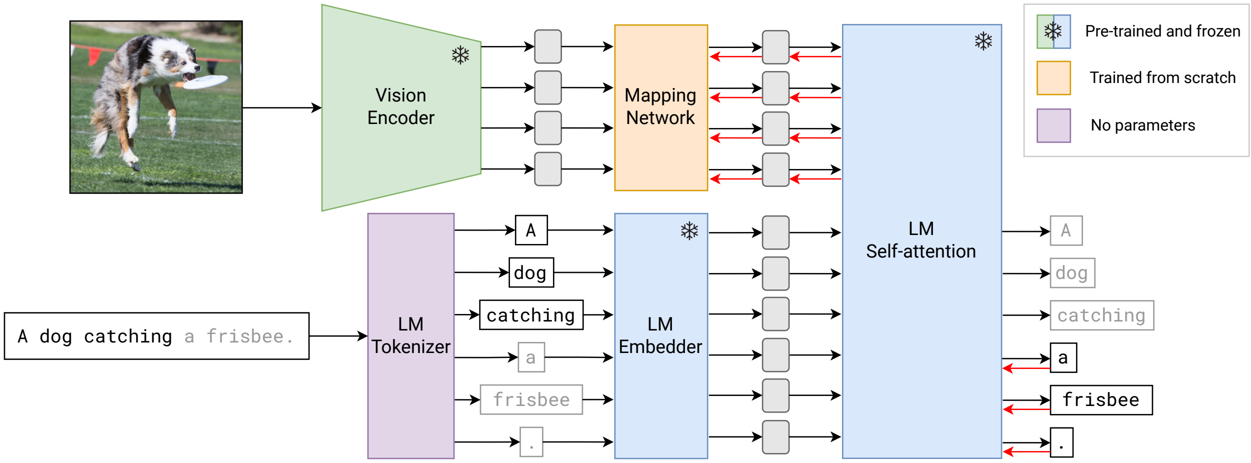


Fig. 2.1. MAPL leverages a pre-trained vision encoder and a pre-trained LM, and learns a small *mapping network* to convert visual features into token embeddings. During training, only the mapping network is updated, keeping the vision encoder and the LM frozen (red arrows indicate gradient flow). At inference time, the system can take as input an arbitrary sequence of interleaved images and text, and generates free-form text as output.

Mapping networks. MAPL trains a mapping network to align the visual and textual representations of the visual encoder and the LM, respectively. The architecture of our mapping network has some similarities with that in ClipCap [**67**] and the Perceiver Resampler in Flamingo [**5**]. They all share a core transformer stack and a fixed number of learned constant embeddings. However, MAPL's mapping network is specifically designed to be parameter-efficient while maintaining expressivity (Sec. 2.5.1), containing only 3.4M parameters – orders of magnitude fewer than ClipCap's (43M) and Flamingo's (194M).

2.5. Method

MAPL is a vision-language (VL) multimodal model capable of generating text from a combination of visual and textual inputs. Our model builds on top of pre-trained vision-only and language-only models and leverages their strong generalization capabilities (e.g., zero-shot transfer, in-context learning) to tackle unseen VL tasks. MAPL is agnostic to the choice of these pre-trained unimodal models as long as they show such capabilities (Sec. 2.6.5). Concretely, MAPL maps the image representations from a vision encoder's output embedding space to a LM's token embedding space, so that the LM can be conditioned both on visual and textual information. To this end, we train a *mapping network* with an image captioning objective (Sec. 2.5.1, 2.5.2), while keeping the weights of the vision encoder and the LM frozen. Once the mapping network is trained, MAPL can be prompted with a few examples of unseen VL tasks and predict the response via text generation (Sec. 2.5.3). The overall model architecture is depicted in Figure 2.1.

2.5.1. Architecture

Pre-trained vision encoder. The vision encoder extracts a compact representation from an image. We use a CLIP [**70**] pre-trained vision encoder, which is trained on web-scale data and has shown strong zero-shot transfer capabilities to unseen image domains. In particular, we use CLIP's ViT-L/14 backbone [**26**] since we empirically found it yields the best downstream VL performance among all variants. We use the flattened grid of spatial features (16×16) before the final projection layer and the representation corresponding to the `[class]` token, resulting in a sequence of $L_i = 257$ vectors of dimensionality $D_i = 1024$ each. This sequence of vectors is then fed to the mapping network.

Pre-trained autoregressive language model. Given an input text, the language model (LM) predicts its most likely completion by generating free-form text. For our LM, we use a pre-trained GPT-J model [**99**] [1], a publicly-released 6B-parameter autoregressive LM trained on the Pile dataset [**31**]. We chose this LM due to its strong in-context learning abilities, similar to that of GPT-3 [**11**] (which is not publicly available). The LM takes as input a text

[1] We also experiment with an OPT model, see Sec. 2.6.5.

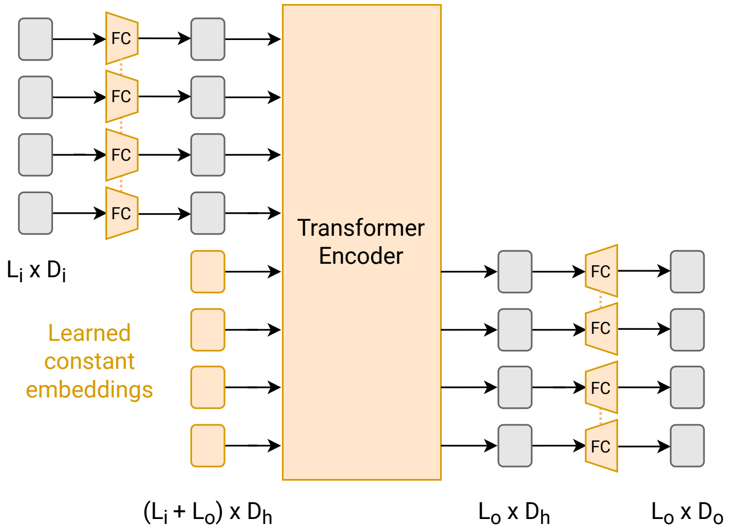


Fig. 2.2. The mapping network takes a flattened grid of L_i visual features of dimension D_i each from the vision encoder and transforms it into a sequence of token embeddings of length L_o and dimension D_o, where D_o is the token embedding dimension of the LM. Note that the parameters are shared across fully-connected (FC) layers, on both sides of the encoder transformer.

string, which is first divided into a sequence of discrete tokens by the LM's tokenizer. Each token is then individually transformed into a continuous embedding (of size $D_o = 4096$) by the LM's embedder. The sequence of token embeddings is fed to the self-attention layers in the LM's transformer block (using causal attention), which outputs a sequence of categorical distributions over the token vocabulary. Finally, a decoding mechanism generates free-form text from these distributions (greedy decoding in our case).

Mapping network. The mapping network transforms a sequence of visual features from the vision encoder to a sequence of continuous embeddings which can be consumed by the LM's transformer. We design our mapping network considering the trade-off between expressivity (to learn a good mapping) and parameter count. Our architecture is based on a transformer encoder with 4 layers and 8 heads each. This transformer could directly take a sequence of projected visual features (from D_i to D_o) and output a sequence of embeddings of size D_o. However, in order to keep a low parameter count, we decouple the transformer hidden size D_h from the visual feature size D_i and the LM embedding size D_o by introducing a dimensionality bottleneck (Figure 2.2). In particular, each visual feature is first linearly projected from $D_i = 1024$ to $D_h = 256$ using a set of fully-connected (FC) layers. This sequence of projected features is then fed to the transformer, and the output representations are linearly projected from D_h to $D_o = 4096$ using another set of FC layers. To further reduce the parameter count of our mapping network, we share parameters across all FC layers in each set.

Yet another idea we use in our mapping network is to decouple the output sequence length of the transformer (L_o) from the input sequence length (L_i). We do this to obtain a much smaller $L_o = 32$ compared to $L_i = 257$, in order to reduce the computational complexity in the subsequent LM's self-attention layers, which in turn speeds up training and inference time. To achieve this decoupling, inspired by DETR [**13**], we concatenate a small and fixed number (L_o) of learned constant embeddings with the input sequence of the transformer and only use the output representations corresponding to these constant embeddings (Figure 2.2). Note that these output representations are conditioned on the input visual features via cross-attention in the transformer. The resulting mapping network architecture is shown in Figure 2.2. In total, our mapping network contains only 3.4M parameters. Since this is the only trainable component of our model, MAPL has orders of magnitude fewer total trainable parameters than existing methods such as Frozen (40.3M) or Flamingo (10.2B).

2.5.2. Training

Following previous works [**94, 29**], we train our model using a standard language modeling objective on image captions with teacher forcing [**49**], i.e., we minimize the negative log-likelihood of the reference captions under the LM conditioned on the corresponding images. We only train the mapping network (from scratch) while keeping the vision encoder and the LM entirely frozen. This preserves the pre-trained models' capabilities while making the system modular and parameter-efficient. Even though the LM's weights are kept frozen, gradients are still back-propagated through its self-attention layers to train the mapping network.

2.5.3. Zero- and Few-shot Evaluation

Once the mapping network is trained, MAPL can tackle unseen VL tasks by prompting the LM with a combination of visual and textual inputs. We study zero-shot transfer to unseen image captioning benchmarks and few-shot transfer (via in-context learning) to the unseen task of visual question answering (VQA). For image captioning, we simply feed the mapped image embedding to the LM and start generating a caption. For zero-shot VQA, following [**94**], we feed the mapped image embedding followed by the text `"Please answer the question. Question: {question} Answer:"`[2] and start generating the answer. For n-shot VQA, we select n support examples ($image, question, answer$) from the training set at random, and prepend them to the query; for each support example, we concatenate

[2] Here `{question}` indicates a placeholder which gets replaced by the corresponding question in each example. Same applies to `{answer}` in the few-shot setting.

the mapped image embedding with the text `"Please answer the question. Question: {question} Answer: {answer}"`.

2.6. Experiments

2.6.1. Experimental settings

Evaluation benchmarks. We evaluate MAPL on several VL benchmarks spanning VQA and image captioning. Note that our model is never trained for the task of VQA. For VQA, we evaluate on the validation splits of VQAv2 [**32**], OK-VQA [**64**], TextVQA [**86**] and VizWiz-VQA [**33**], and report performance using VQA accuracy (after the standard normalization [**7**]). For image captioning, we evaluate on the Karpathy-test split [**44**] of COCO Captions [**17**], and the validation splits of Conceptual Captions (CC) [**81**][3], TextCaps [**84**] and VizWiz-Captions [**33**], and report performance using the BLEU@4, ROUGE-L, METEOR, CIDEr and SPICE metrics.

Training settings. We consider two settings to train our mapping network: domain-agnostic and in-domain training (described below). For each of these settings, we also study low-data learning by training our model on randomly sampled subsets of 1% training image-text pairs. Such low-data learning is useful when it is difficult to train models on large-scale data due to constraints on compute resources, data availability, etc.

For **domain-agnostic training**, we use the CC dataset, which is gathered by automatically scraping images and their corresponding alt-text fields from web pages. Thus, this dataset is not as clean as manually-curated datasets such as COCO Captions (e.g., the caption may not describe the image). Nevertheless, due to its large size (3.3M) and great diversity, it is the most commonly used dataset for domain-agnostic pre-training of VL models. However, for our model – having orders of magnitude less trainable parameters than other methods –, we observed the negative effect of noise in CC to be stronger than the positive effect of its large size (Sec. 2.6.4). Therefore, we train MAPL on a filtered version of CC (CC-clean) consisting of the top 398K most similar image-text pairs ranked by CLIP's image-text similarity score.[4] For completeness, we also report MAPL's performance when trained on the unfiltered CC dataset.

For **in-domain training**, we use image-caption pairs that come from the same domain as the downstream task domains, i.e., they have similar image and language distributions as those in the downstream datasets. For the image captioning downstream task, this amounts to the IID setting. The in-domain image captioning and VQA dataset pairs we consider are shown in Table 2.1. Each pair uses the same set of images, and focuses on the same set of VL

[3] Due to broken image URLs, we only managed to download 13K out of 15K validation images.

[4] We selected a threshold on CLIP's similarity score such that the size of the filtered dataset is comparable to the size of manually curated datasets such as COCO Captions.

skills; for instance, scene understanding (COCO Caps and VQAv2), reading and reasoning about text in images (TextCaps and TextVQA), understanding images captured by visually-impaired users (VizWiz-Caps and VizWiz-VQA), thus leading to similar image and language distributions across image-captioning and VQA. We train MAPL on both 100% and 1% of in-domain image-caption data and evaluate on all downstream benchmarks (including out-of-domain ones, e.g., VizWiz-VQA when trained on COCO Caps). Such in-domain training can be useful when it is difficult to first train on large-scale domain-agnostic data and then adapt to in-domain data by either fine-tuning or few-shot prompting.

	VQAv2	**OK-VQA**	**TextVQA**	**VizWiz-VQA**
COCO Caps	✓	✓		
TextCaps			✓	
VizWiz-Caps				✓

Table 2.1. In-domain dataset pairs.

Training details. For Conceptual Captions, TextCaps and VizWiz-Captions, we carve out a minival split consisting of 6% of training examples and train on the remaining 94%; for COCO Captions, we use the Karpathy-val split as minival. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.01. The learning rate is increased linearly from 0 to 3×10^{-4} (7×10^{-4} for OPT-based models) over the first 1500 steps (15 for 1% of data) and kept constant for the rest of training. We use a batch size of 128 and we do early stopping based on the minival loss. We do not add any special tokens at the beginning of sentence, as GPT-J was not trained with `<BOS>` tokens. In order to fit a 6B-parameter LM into GPU memory, we use DeepSpeed ZeRO [**74**] stage 2 optimizations. Freezing the LM's weights also brings massive savings in GPU memory during training, as fine-tuning with an Adam-based optimizer would require at least 4× GPU memory to store gradients, average, and squared average of the gradients. The whole system was trained on 4 A100 (40GB) GPUs for about 4 hours (for the CC-clean dataset). Unless otherwise stated, we repeat the experiments with two different random seeds and report the average performance.

Existing methods and baselines. We report the performance of several baselines and existing methods. First, to verify that the LM in MAPL is not ignoring the visual input, inspired by [**94**], we train a blind version of MAPL (MAPLblind) where the input images are replaced with zeros but the mapping network weights are still trained (to serve as prompt-tuning for the LM). Second, to estimate the upper-bound on how well we can do in VQA by representing images with text (rather than with continuous embeddings), we evaluate PICa [**104**], which directly prompts the LM with image captions, followed by questions for VQA. We reimplement PICa (denoted PICa*) using MAPL's LM (and evaluate on VQAv2 and OK-VQA using ground-truth COCO captions) for controlled comparison. Third, we compare MAPL with Frozen [**94**], as this is the most similar method to ours that also uses a frozen LM. We reimplement Frozen (denoted Frozen*) using MAPL's LM for controlled

	Trainable	**Training**	**n-shot VQAv2**			**n-shot OK-VQA**			**n-shot TextVQA**			**n-shot VizWiz-VQA**			**n-shot Overall**		
	params	**examples**	**0**	**4**	**8**	**0**	**4**	**8**	**0**	**4**	**8**	**0**	**4**	**8**	**0**	**4**	**8**
Existing methods using domain-agnostic training																	
Frozen	40.3M†	3.3M	29.50	38.20	-	5.90	12.60	-	-	-	-	-	-	-	-	-	-
MAGMA $_{\text{CC12M}}$	243M†	3.8M	36.90	45.40	-	13.90	23.40	-	-	-	-	5.60	10.60	-	-	-	-
VLKD $_{\text{CC3M}}$	406M	3.3M	38.60	-	-	10.50	-	-	-	-	-	-	-	-	-	-	-
LiMBeR-CLIP‡	12.6M†	3.3M	33.33	40.34	-	-	-	-	-	-	-	-	-	-	-	-	-
Flamingo‡	10.2B	>2.1B	-	-	-	50.60	57.40	57.50	35.00	36.50	37.30	-	-	-	-	-	-
100% domain-agnostic training																	
MAPL-blind $_{\text{CC-clean}}$	3.4M	374K	20.62	35.01	35.11	4.84	14.68	14.28	3.68	5.43	5.82	3.18	8.65	9.55	8.08	15.94	16.19
Frozen* $_{\text{CC-clean}}$	40.3M	374K	25.98	37.80	38.52	5.51	18.86	19.91	5.11	6.15	6.30	4.33	11.28	16.68	10.23	18.52	20.35
MAPL $_{\text{CC-clean}}$	3.4M	374K	**33.54**	**45.13**	**45.21**	**13.84**	**24.25**	**23.93**	**8.26**	**8.88**	**8.77**	**11.72**	**18.46**	**19.52**	**16.84**	**24.18**	**24.36**
1% domain-agnostic training																	
Frozen* $_{\text{CC-clean}}$	40.3M	3.7K	26.22	36.69	37.41	5.50	**18.76**	**20.51**	5.71	**7.19**	7.53	3.83	**11.71**	**16.66**	10.31	**18.58**	**20.53**
MAPL $_{\text{CC-clean}}$	3.4M	3.7K	**30.80**	**37.38**	**37.95**	**8.77**	18.18	19.15	**6.40**	7.07	**7.74**	**5.68**	9.26	10.58	**12.91**	17.97	18.85
100% in-domain training																	
PICa*	0	0	20.61	46.86	47.80	11.84	**31.28**	**33.07**	-	-	-	-	-	-	-	-	-
Frozen* $_{\text{COCO}}$	40.3M	414K	32.09	38.90	39.42	9.81	20.72	21.83	7.54	6.82	6.74	5.87	12.07	17.35	13.82	19.63	21.33
Frozen* $_{\text{TextCaps}}$	40.3M	103K	32.49	37.39	38.03	11.34	19.87	20.82	8.83	7.33	7.51	6.25	12.26	16.86	14.73	19.21	20.80
Frozen* $_{\text{VizWiz}}$	40.3M	110K	26.93	37.38	37.91	5.85	19.12	20.64	6.38	7.44	7.47	5.57	13.06	18.06	11.18	19.25	21.02
MAPL $_{\text{COCO}}$	3.4M	414K	**43.51**	**48.75**	**48.44**	**18.27**	31.13	31.63	10.99	11.10	11.08	**14.05**	17.72	19.18	21.70	**27.17**	**27.58**
MAPL $_{\text{TextCaps}}$	3.4M	103K	38.83	43.34	43.43	16.33	25.07	25.92	**22.27**	**19.53**	**19.75**	12.31	16.69	18.18	**22.43**	26.15	26.82
MAPL $_{\text{VizWiz}}$	3.4M	110K	32.80	42.94	43.20	11.70	24.91	25.73	9.27	10.36	10.23	10.42	**20.63**	**23.10**	16.05	24.71	25.56
1% in-domain training																	
Frozen* $_{\text{COCO}}$	40.3M	4.1K	30.18	37.23	37.89	9.33	19.60	20.71	7.43	7.65	7.67	4.37	12.00	16.48	12.83	19.12	20.69
Frozen* $_{\text{TextCaps}}$	40.3M	1.0K	32.09	36.72	37.25	10.75	18.85	19.51	8.17	7.57	7.28	5.39	11.79	16.20	14.10	18.73	20.06
Frozen* $_{\text{VizWiz}}$	40.3M	1.1K	29.62	37.30	37.87	7.57	19.36	20.60	7.16	7.17	7.25	4.53	**12.51**	**17.56**	12.22	19.08	**20.82**
MAPL $_{\text{COCO}}$	3.4M	4.1K	**37.69**	**40.42**	**40.84**	**13.92**	**21.66**	**22.41**	8.30	6.96	6.84	**6.94**	10.72	12.43	**16.71**	**19.94**	20.63
MAPL $_{\text{TextCaps}}$	3.4M	1.0K	33.57	36.70	36.87	12.46	17.45	18.21	**9.34**	**8.29**	**8.62**	6.54	9.58	11.62	15.48	18.00	18.83
MAPL $_{\text{VizWiz}}$	3.4M	1.1K	31.88	36.81	37.04	9.59	17.64	17.64	7.25	5.99	6.04	4.73	9.48	11.33	13.36	17.48	18.01

Table 2.2. Evaluation on few-shot VQA. For MAGMA $_{\text{CC12M}}$ and VLKD $_{\text{CC3M}}$, we report their best results when training only on domain-agnostic data (CC12M and CC3M, respectively). (†) indicates our informed estimation. (‡) indicates concurrent work.

comparison. Lastly, we report the performance of other methods similar to MAPL: MAGMA [**29**], VLKD [**23**], LiMBeR [**66**], ClipCap [**67**] and the published numbers from Frozen [**94**].[5] Note that all these methods (unless otherwise noted) are trained on domain-agnostic data, so we only compare with MAPL trained on CC-clean. For completeness, we also report results from Flamingo [**5**], which has orders of magnitude more learnable parameters than MAPL and is trained on considerably more data.

2.6.2. Evaluation of domain-agnostic learning

We report few-shot VQA results in Table 2.2 and image captioning results in Table 2.3. Subscripts in the first column denote the training dataset. *Overall* accuracies denote average of per-benchmark accuracies. First, we see that MAPL$_{\text{CC-clean}}$ substantially outperforms MAPLblind$_{\text{CC-clean}}$ both on VQA and image captioning, proving that the visual inputs are not ignored by the LM in MAPL. Second, we find that MAPL$_{\text{CC-clean}}$ outperforms Frozen* $_{\text{CC-clean}}$ by a considerable margin on all VL benchmarks (with overall accuracy improvements of +6.61% 0-shot and +5.66% 4-shot on VQA tasks, +4.35 BLEU@4 and +33.55 CIDEr on image captioning tasks). Importantly, this is achieved while training an order of magnitude fewer parameters (3.4M vs 40.3M). Next, MAPL$_{\text{CC-clean}}$ is competitive compared to existing methods (MAGMA, VLKD, ClipCap) and concurrent work LiMBeR, despite training one-two orders of magnitude fewer parameters on significantly less multimodal data. Lastly,

[5] We only add results which are reported on the same dataset splits as in MAPL.

	Trainable	**Training**	**CC**		**COCO**		**TextCaps**		**VizWiz-Caps**		**Overall**	
	params	**examples**	**B@4**	**CIDEr**	**B@4**	**CIDEr**	**B@4**	**CIDEr**	**B@4**	**CIDEr**	**B@4**	**CIDEr**
Existing methods using domain-agnostic training												
ClipCap $_{\text{CC3M}}$	43M	3.3M	-	71.82	-	-	-	-	-	-	-	-
VLKD $_{\text{CC3M}}$	406M	3.3M	-	-	18.20	61.10	-	-	-	-	-	-
100% domain-agnostic training												
MAPL-blind $_{\text{CC-clean}}$	3.4M	374K	0.35	5.05	2.75	5.75	1.35	2.15	1.50	1.80	1.49	3.69
Frozen* $_{\text{CC-clean}}$	40.3M	374K	2.45	22.60	5.25	13.90	2.65	4.60	2.05	2.65	3.10	10.94
MAPL $_{\text{CC-clean}}$	3.4M	374K	**6.75**	**79.75**	**12.30**	**54.30**	**5.80**	**22.95**	**4.95**	**20.95**	**7.45**	**44.49**
1% domain-agnostic training												
Frozen* $_{\text{CC-clean}}$	40.3M	3.7K	0.75	6.55	3.05	5.25	1.70	1.65	1.50	1.40	1.75	3.71
MAPL $_{\text{CC-clean}}$	3.4M	3.7K	**1.75**	**19.65**	**5.80**	**17.85**	**2.70**	**5.40**	**2.15**	**4.85**	**3.10**	**11.94**
100% in-domain training												
Frozen* $_{\text{COCO}}$	40.3M	414K	0.65	9.05	20.05	61.35	6.95	11.75	5.45	6.20	8.28	22.09
Frozen* $_{\text{TextCaps}}$	40.3M	103K	0.20	3.55	4.05	6.70	8.85	16.95	4.40	5.25	4.38	8.11
Frozen* $_{\text{VizWiz}}$	40.3M	110K	0.25	4.40	3.75	6.05	4.10	5.65	19.00	76.85	6.78	23.24
ClipCap $_{\text{COCO}}$	43M	414K	-	-	33.53	113.08	-	-	-	-	-	-
MAPL $_{\text{COCO}}$	3.4M	414K	**2.25**	**34.50**	**36.45**	**125.20**	16.60	41.40	18.00	41.35	**18.33**	**60.61**
MAPL $_{\text{TextCaps}}$	3.4M	103K	0.90	13.05	9.80	28.65	**18.35**	**62.55**	11.20	31.85	10.06	34.03
MAPL $_{\text{VizWiz}}$	3.4M	110K	0.90	18.80	13.55	48.35	11.35	31.20	**34.70**	**141.30**	15.13	59.91
1% in-domain training												
Frozen* $_{\text{COCO}}$	40.3M	4.1K	0.25	3.60	6.20	12.80	2.80	3.15	2.85	2.30	3.03	5.46
Frozen* $_{\text{TextCaps}}$	40.3M	1.0K	0.10	2.60	1.65	2.80	3.65	5.00	2.00	2.25	1.85	3.16
Frozen* $_{\text{VizWiz}}$	40.3M	1.1K	0.20	3.40	2.90	3.20	3.35	3.45	12.70	40.55	4.79	12.65
MAPL $_{\text{COCO}}$	3.4M	4.1K	**0.80**	**12.10**	**19.65**	**65.90**	7.00	12.85	6.20	9.60	**8.41**	**25.11**
MAPL $_{\text{TextCaps}}$	3.4M	1.0K	0.30	3.90	4.10	8.05	**8.35**	**16.90**	5.00	7.25	4.44	9.03
MAPL $_{\text{VizWiz}}$	3.4M	1.1K	0.20	3.90	2.95	4.80	3.45	5.05	**18.40**	**71.10**	6.25	21.21

Table 2.3. Evaluation on image captioning. For VLKD $_{\text{CC3M}}$, we report their best results when training only on domain-agnostic data (CC3M).

	Training	**VQAv2**	**OK-VQA**	**TextVQA**	**VizWiz-VQA**	**CC**	**COCO**	**TextCaps**	**VizWiz-Caps**	**Overall**	
	examples	**4-shot**	**4-shot**	**4-shot**	**4-shot**	**CIDEr**	**CIDEr**	**CIDEr**	**CIDEr**	**4-shot**	**CIDEr**
Frozen* $_{\text{CC-clean}}$	0.4M	37.79	**19.29**	**6.25**	**11.11**	22.70	14.00	5.00	2.70	**18.61**	11.10
Frozen* $_{\text{CC-cleanish}}$	1.0M	**37.82**	18.49	6.12	10.16	37.60	20.60	6.60	3.20	18.15	17.00
Frozen* $_{\text{CC}}$	2.7M	37.81	18.33	5.56	9.97	**57.60**	**22.20**	**8.00**	**4.20**	17.92	**23.00**
MAPL $_{\text{CC-clean}}$	0.4M	44.35	24.03	**9.65**	17.33	72.70	**54.60**	**23.80**	**21.10**	23.84	43.05
MAPL $_{\text{CC-cleanish}}$	1.0M	**46.63**	**25.99**	8.48	**19.65**	88.30	54.10	22.30	19.80	**25.19**	**46.13**
MAPL $_{\text{CC}}$	2.7M	43.26	20.96	5.20	19.31	**101.10**	44.10	16.70	15.90	22.18	44.45

Table 2.4. Impact of data quality and size. These experiments are run with one seed only.

MAPL$_{\text{CC-clean}}$'s performance is still far from the performance of Flamingo, which trains orders of magnitude more parameters on orders of magnitude more data. However, we believe MAPL to be an effective method for scenarios with constrained computational resources. For MAPL's qualitative results, see App. 2.6.6.

Low-data learning. When trained on only 1% domain-agnostic data, MAPL$_{\text{CC-clean}}$ outperforms Frozen* $_{\text{CC-clean}}$ for all image captioning evaluations (by +1.35 BLEU@4 and +8.23 CIDEr, overall) and all 0-shot VQA evaluations (by +2.60% overall accuracy), while achieving competitive performance on 4- and 8-shot VQA evaluations. In summary, these results show the effectiveness of our method in low-data settings, highlighting its usefulness for applications where data is scarce.

	Vision	Language	Mapping	VQAv2		OK-VQA		TextVQA		VizWiz-VQA		CC	COCO	TextCaps	VizWiz-Caps	Overall		
	encoder	**model**	**network**	**0-shot**	**4-shot**	**0-shot**	**4-shot**	**0-shot**	**4-shot**	**0-shot**	**4-shot**	**CIDEr**	**CIDEr**	**CIDEr**	**CIDEr**	**0-shot**	**4-shot**	**CIDEr**
Frozen*	NF-ResNet-50	GPT-J	Transformer	27.98	36.66	5.88	18.44	4.56	7.87	3.67	10.32	21.79	14.87	5.42	3.03	10.52	18.32	11.28
Frozen*	ViT-L/16	GPT-J	Transformer	27.82	36.60	5.64	16.26	3.67	5.27	4.70	11.77	13.19	8.77	2.88	2.35	10.46	17.48	6.80
Frozen*	ViT-L/14	GPT-J	Transformer	24.45	36.03	4.14	16.27	3.91	5.33	3.27	10.09	13.89	8.27	2.66	2.50	8.94	16.93	6.83
MAPL	CLIP-ViT-L/14	GPT-J	Transformer	**33.54**	**45.13**	13.84	24.25	8.26	**8.88**	11.72	**18.46**	**79.75**	54.30	22.95	**20.95**	**16.84**	**24.18**	**44.49**
MAPL	IN-NF-ResNet-50	GPT-J	Transformer	28.12	40.86	10.86	21.66	6.15	7.01	6.40	14.22	39.64	32.99	12.41	9.55	12.88	20.94	23.65
MAPL	IN-ViT-L/16	GPT-J	Transformer	31.70	43.75	11.13	**25.50**	6.16	7.40	8.93	16.45	56.33	45.80	17.12	16.28	14.48	23.28	33.88
Frozen*	NF-ResNet-50	OPT-6.7B	Transformer	30.16	32.72	8.10	13.79	5.44	6.81	7.15	6.77	25.40	17.80	6.90	4.00	12.71	15.02	13.53
MAPL	CLIP-ViT-L/14	OPT-6.7B	Transformer	23.26	33.95	**15.27**	16.25	**8.90**	6.41	**15.40**	9.47	66.60	**54.40**	**23.30**	19.60	15.71	16.52	40.98
MAPL	CLIP-ViT-L/14	GPT-J	Linear	30.55	37.09	12.20	16.69	7.02	5.80	8.81	12.49	60.00	43.80	18.30	13.70	14.65	18.02	33.95
MAPL	CLIP-ViT-L/14	GPT-J	MLP	28.99	43.69	11.07	25.33	6.60	8.39	9.73	17.14	70.40	49.10	20.90	20.30	14.10	23.64	40.18

Table 2.5. Ablation studies. We assess the impact of the choice of vision encoder (top), LM (middle) and mapping network architecture (bottom). All models are trained on 100% of CC-clean with a single seed. IN stands for ImageNet pre-training.

2.6.3. Evaluation of in-domain learning

In Tables 2.2 and 2.3, we observe that both MAPL and Frozen* benefit from directly training on in-domain data, compared to few-shot transfer from large-scale domain-agnostic pretraining. For instance, MAPL$_{\text{COCO}}$ and Frozen* $_{\text{COCO}}$ respectively outperform MAPL$_{\text{CC-clean}}$ and Frozen* $_{\text{CC-clean}}$ on VQAv2, OK-VQA and COCO Captions when trained on 100% of data. Interestingly, this performance gap is larger for MAPL compared to Frozen* by +2% 0-shot accuracy and +3.77% 4-shot accuracy averaged across VQAv2 and OK-VQA, and +9.35 BLEU@4 and +23.45 CIDEr on COCO Captions. A similar trend can be observed for TextCaps and TextVQA. Surprisingly, for 0-shot VQA and image captioning, training on just 1% of in-domain data outperforms 100% CC training for all benchmarks (except VizWiz-VQA) and both models. These results demonstrate the benefits of in-domain learning. When comparing MAPL vs. Frozen*, we observe that MAPL outperforms Frozen* for all tasks and benchmarks (except VizWiz-VQA) under both 100% and 1% in-domain settings. In fact, MAPL trained on just 1% in-domain data outperforms Frozen* trained on 100% in-domain data by +3.41% 0-shot accuracy and +1.14% 4-shot accuracy averaged across VQAv2, OK-VQA and TextVQA. Thus, MAPL is more effective than Frozen* at in-domain learning.

Contrary to the above trends, we observe that MAPL$_{\text{VizWiz}}$ under 1% in-domain training performs worse than MAPL $_{\text{COCO}}$ or MAPL $_{\text{TextCaps}}$ when evaluated on VizWiz-VQA. We hypothesize the visual embeddings extracted from CLIP's vision encoder for VizWiz images are not as good as those for COCO or TextCaps' images because the distribution of images in VizWiz (captured by visually-impaired people) is rather different from the distribution of images CLIP is trained on (scraped from the web), whereas for COCO and TextCaps this isn't the case. When training MAPL's mapping network on only 1% of VizWiz data, we believe the data is not large enough to compensate for the OOD pretrained vision encoder, so MAPL trained on COCO/TextCaps performs better on VizWiz-VQA. For in-domain training with 100% of data and 4/8-shot VQA, the mapping network has enough data to learn from and compensate for the OOD phenomenon. On the other hand, Frozen* does not suffer from this issue because its vision encoder is trained from scratch, allowing it to adapt to the image distribution.

Lastly, we observe that $\text{MAPL}_{\text{COCO}}$ outperforms ClipCap $_{\text{COCO}}$ by +2.92 BLEU@4 and +12.12 CIDEr on COCO Captions. $\text{MAPL}_{\text{COCO}}$ also outperforms PICa* (which represents images with ground-truth COCO captions) on VQAv2 and 0-shot OK-VQA, and achieves competitive results on few-shot OK-VQA; this demonstrates representing images with continuous embeddings is beneficial over caption-based image representations. Overall, we see that in-domain learning is beneficial and MAPL is more effective at it than similar methods.

2.6.4. Impact of data quality and size

To measure the impact of noise in the training data, we additionally train MAPL and Frozen* on the *full* CC dataset, consisting of 2.8M[6] examples, as well as on a *clean-ish* version consisting of the 1.0M most similar image-text pairs. In Table 2.4, we observe Frozen* achieves similar performance on few-shot VQA tasks when trained on noisy vs. clean data; however, Frozen*'s performance on image captioning decreases when trained on cleaner but smaller data. In contrast, MAPL generally benefits from cleaner training data, with the exception of evaluation on CC. We hypothesize both models perform better on CC when trained on larger (yet noisier) data because the CC validation set is IID with the full (noisy) CC training set. In the case of Frozen*, as we move away from the IID setting, the benefits from more data start diminishing (CC captioning > other captioning tasks > VQA tasks). For MAPL, the benefit from reduced noise in training data exceeds the degradation caused by a smaller data size, thanks to the reduced number of trainable parameters. These trends align with previous observations that larger models are more robust to noisy training data since they have enough capacity to model both noise and the desired function [**77**], while smaller models are more sample-efficient [**96**], i.e. they need less (clean) data to train effectively. Note that although MAPL's overall performance is higher when training on 1.0M than on 0.4M examples, we decided to train with 0.4M examples because training on ∼2.5× more data (1.0M instead of 0.4M) required ∼5× more iterations (always early-stopping based on validation loss). So we did not think the slight performance increase due to more data was worth the ∼5× longer training time, especially because we were operating under a limited compute budget.

2.6.5. Ablation studies

In this section, we evaluate how the choice of vision encoder, LM and mapping network architecture impact MAPL's performance, and compare it with corresponding versions of

[6] This is not the full 3.3M CC due to broken URLs.

Frozen* (where applicable). Results are presented in Table 2.5. Please refer to App. A.0.5 for more ablations.

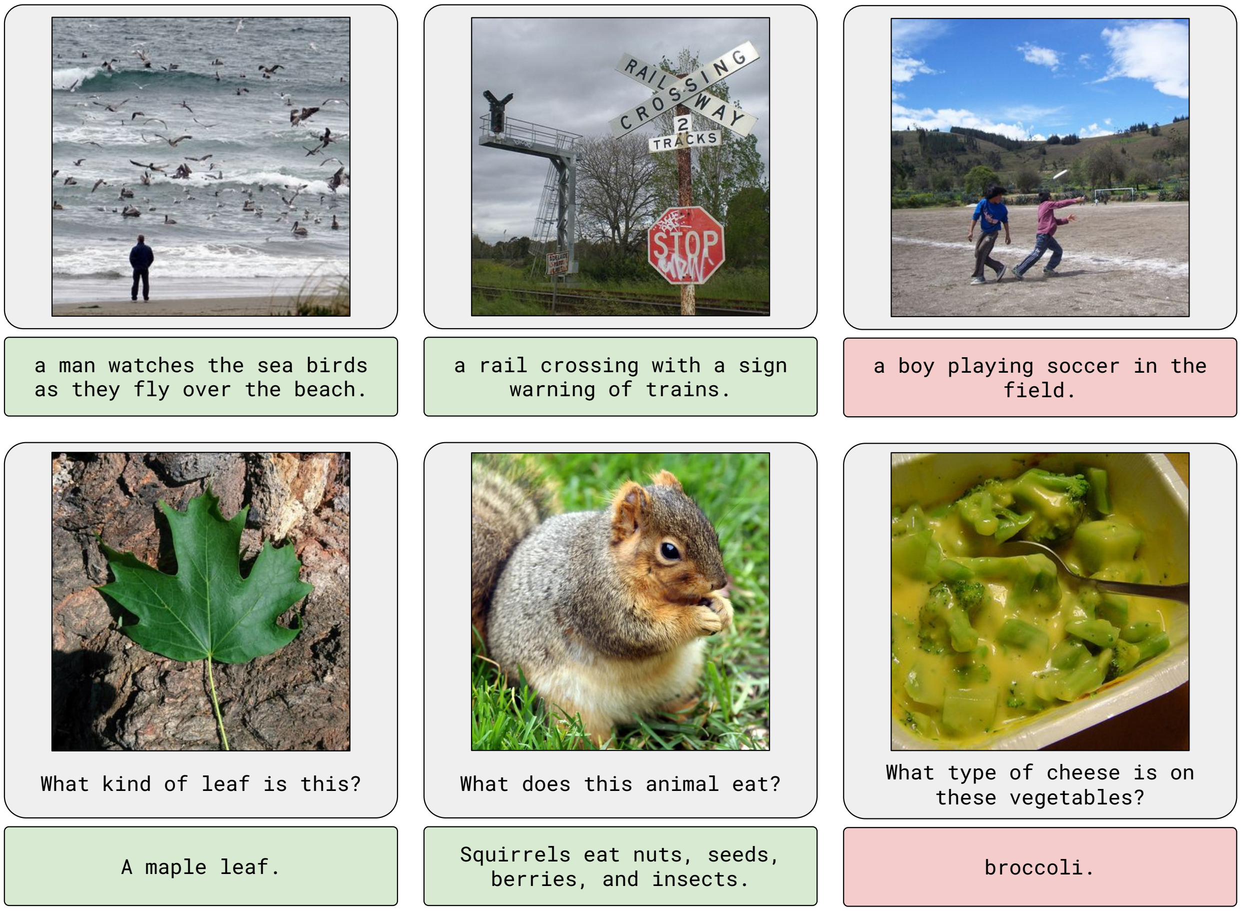
First, to assess the impact of the choice of vision encoder, we train additional versions of MAPL replacing the CLIP pre-trained vision encoder (ViT-L/14 – 303M parameters) with encoders pre-trained on ImageNet: NF-ResNet-50 (23.5M) and ViT-L/16 (303M), and compare their performance with corresponding versions of Frozen*. We observe that: 1) MAPL outperforms Frozen* for each configuration of vision encoder, suggesting that MAPL is **robust to the choice of vision encoder's pre-training data and architecture**; and 2) Frozen* 's performance drops with bigger vision encoders (likely due to more trainable parameters), whereas MAPL improves due to the use of stronger pre-trained encoders. Thus, training the vision encoder from scratch (Frozen*) has limited application, while MAPL's **performance scales alongside the pre-trained vision encoder**.

Next, to evaluate the impact of the choice of LM, we train both MAPL and Frozen* replacing GPT-J by OPT-6.7B [**110**]. We see that in all settings except 0-shot VQAv2, MAPL outperforms Frozen*. See App. A.0.2 for discussion on 0-shot VQAv2 results. This suggests that MAPL is **robust to the choice of LM**. The above results also highlight how MAPL's **modularity** allows to easily replace the pre-trained vision encoder or the LM.

Lastly, to assess the impact of the choice of mapping network architecture, we replace the proposed transformer-based mapping network with two simpler architectures – a linear layer and a 2-layer MLP (see App. A.0.4 for details). We observe both these versions generally underperform the original setting (transformer-based), highlighting the **effectiveness of the proposed design**. We also note that in these simpler versions, the parameter count is directly proportional to the vision encoder's representation size and LM's embedding size, whereas in MAPL we decouple this using a dimensionality bottleneck (Sec. 2.5.1), making our mapping network **more parameter-efficient by design**.

2.6.6. Qualitative results

Figure 2.3 shows some selected samples from the web illustrating our interface at inference time using $\text{MAPL}_{\text{CC-clean}}$. The first two columns show successful results while the last column shows failure cases. For image captioning (top row), success cases show MAPL can generate meaningful and detailed textual descriptions of the scene. For zero-shot VQA (bottom row), success cases indicate that MAPL is able to parse the question and connect visual information to encyclopedic knowledge contained in the pre-trained LM. However, MAPL's visio-linguistic understanding is evidently still far from being perfect. More qualitative results (both success and failure cases) are provided in App. A.0.6.



a man watches the sea birds as they fly over the beach.

a rail crossing with a sign warning of trains.

a boy playing soccer in the field.

What kind of leaf is this?

A maple leaf.

What does this animal eat?

Squirrels eat nuts, seeds, berries, and insects.

What type of cheese is on these vegetables?

broccoli.

Fig. 2.3. Qualitative samples from the web using MAPL$_{\text{CC-clean}}$. (Multimodal) input is in gray, and MAPL's output is in green (success) or red (failure).

2.7. Conclusion

We introduce MAPL, a simple and parameter-efficient method to repurpose pre-trained and frozen unimodal models for multimodal tasks. Our experiments demonstrate that MAPL achieves superior or competitive performance compared to similar methods on several VL benchmarks while training orders of magnitude fewer parameters. Importantly, we also show that MAPL is effective in the low-data and in-domain settings thanks to its reduced number of trainable parameters. We leave as future work exploring training on a weighted mixture of image-text datasets, evaluating on more downstream tasks such as NLVR2 [**88**] and Visual Dialog [**24**], and investigating the use of masked LMs [**80, 22**] with MAPL.

2.8. Limitations

MAPL achieves reasonable performance on VL tasks, but it is still far from the performance of recent methods leveraging large-scale data and compute. On the other hand, MAPL is a preferable alternative in scenarios with constrained computational resources.

We observed our mapping network is sensitive to initialization, so different random seeds can yield non-negligible variance in downstream performance. We think this might be related to the reduced number of trainable parameters. We tried to reduce the effect of this variance by reporting average performance across different seeds. We also observed MAPL struggles to leverage more shots for in-context learning. We hypothesize this could be caused by our model being trained on single image-caption pairs – as opposed to the sequences of multiple images and texts seen during few-shot transfer, so a better pretext task might help (see App. A.0.1 for further discussion).

MAPL builds on top of pre-trained vision-only and language-only models, inheriting their capabilities but also their limitations. An important risk is that our model might inherit the existing social, gender or racial biases of pre-trained models. However, our limited qualitative analysis (see App. A.0.8) shows that providing visual information significantly changes the prior answer distribution of the LM. Therefore, how much of the underlying bias is retained remains an empirical question.

2.9. Ethics Statement

Model recycling. MAPL reuses vision-only and language-only foundation models. Hence, the expensive computational resources used to train these models can be amortized to help reduce energy and carbon costs.

Public datasets. MAPL is trained uniquely on publicly available datasets, which facilitates reproducibility and provides transparency on the origin and the characteristics of the data the model has seen.

Undesired biases. MAPL could be exposed to undesired biases from different sources. The pre-trained vision encoder might have been trained with data where certain races or genders are underrepresented, hence biasing our representation of images. The pre-trained LM might also be biased towards generating toxic or offensive language when fed with certain prompts. Finally, the image-text data used to align the representation spaces of such models was annotated by humans, so it might reflect a biased view of the world.

Broader impact. This work shows how one can easily adapt pre-trained vision encoders and LMs for multimodal tasks. Given the parameter-efficiency of our method, we believe it should be of great interest to the sections of the community that do not have access to large compute resources (e.g., small academic labs and independent researchers), and for low-data applications. While MAPL can be applied in many useful applications (e.g., aiding visually-impaired people), it also makes it simpler to create malicious or offensive multimodal systems from existing unimodal models. Further research efforts are needed on how to safely deploy such systems so that their behavior always aligns with ethical values.

Chapter 3

An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics

3.1. Prologue to Paper

3.1.1. Paper Details

An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics, Saba Ahmadi, Aishwarya Agrawal.

This paper is under review at The European Chapter of the Association for Computational Linguistics (EACL), 2024. This work was also presented at Workshop on Open-Domain Reasoning Under Multi-Modal Settings at IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2023.

3.1.2. My Contributions

I designed the analysis and experiments presented in this paper under the guidance of Professor Aishwarya Agrawal. Furthermore, I executed all aspects of the implementation and experiments.

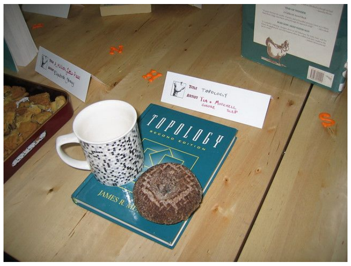
3.2. Abstract

Recently, reference-free metrics such as CLIPScore [**37**], UMIC [**50**], and PAC-S [**79**] have been proposed for automatic reference-free evaluation of image captions. Our focus lies in evaluating the robustness of these metrics in scenarios that require distinguishing between two captions with high lexical overlap but very different meanings. Our findings reveal that despite their high correlation with human judgments, CLIPScore, UMIC, and PAC-S struggle to identify fine-grained errors. While all metrics exhibit strong sensitivity to visual grounding errors, their sensitivity to caption implausibility errors is limited. Furthermore, we found that all metrics are sensitive to variations in the size of image-relevant objects mentioned in the caption, while CLIPScore and PAC-S are also sensitive to the number of mentions of image-relevant objects in the caption. Regarding linguistic aspects of a caption, all metrics show weak comprehension of negation, and CLIPScore and PAC-S are insensitive to the structure of the caption to a great extent. We hope our findings will guide further improvements in reference-free evaluation of image captioning.

3.3. Introduction

Image caption quality has been traditionally evaluated using a reference-based approach, with metrics like BLEU [**68**], ROUGE [**57**], METEOR [**8**], and CIDEr [**98**] assessing the lexical overlap between generated and reference captions. However, this approach is restrictive as the set of references may not capture the full range of valid captions, and furthermore, lexical overlap-based metrics tend to favor captions with similar vocabulary but different meanings. To address these limitations, recent studies like CLIPScore [**37**], UMIC [**50**] and PAC-S [**79**] have proposed reference-free approaches for evaluating image caption quality, which more closely aligns with human judgments. These approaches leverage large pre-trained image-text matching models to measure the similarity between a given image and a candidate caption. However, the evaluation benchmarks for these metrics do not necessarily involve differentiating between captions with significant lexical overlap but vastly different meanings (Fig. 3.1). In this work, we evaluate the robustness of these reference-free metrics in scenarios where the correct and incorrect captions have high lexical overlap. To our surprise, we found that **all metrics fail to distinguish between correct and incorrect captions ∼46% of the time**.

In a pursuit to identify what aspects of a caption (e.g., plausibility, visual grounding, number and size of objects mentioned in the caption, negation and sentence structure) these metrics are most sensitive to, we conduct several controlled experiments, varying one aspect at a time. We found that:



Candidate Captions	CLIPScore	UMIC	PAC-S
The title of the book is topology.	0.472	0.347	0.555
The title of the book is muffin.	0.546	0.446	0.623

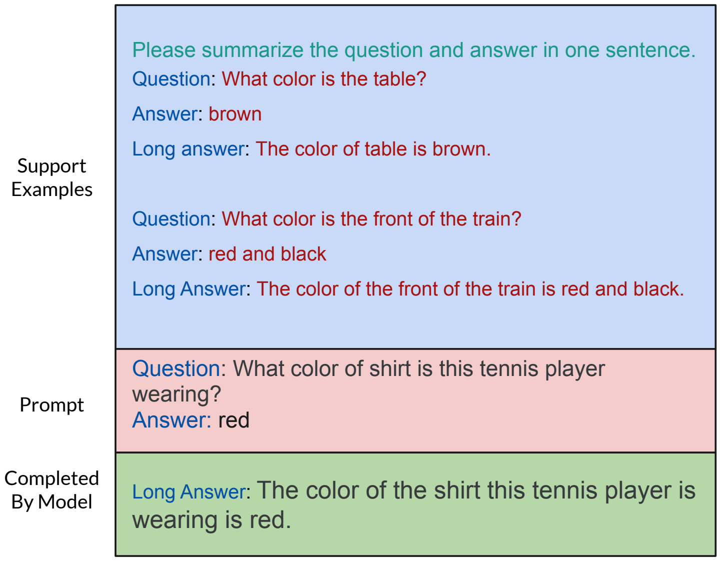
Fig. 3.1. Recently proposed reference-free image captioning evaluation metrics such as CLIPScore, UMIC, and PAC-S are far from perfect. This figure shows how these metrics cannot tell apart an incorrect caption (shown in red) from a correct caption when there is a high lexical overlap between them.

- All metrics show limited sensitivity to caption implausibility errors but a heightened sensitivity to visual grounding errors.
- CLIPScore and PAC-S show high sensitivity to the number of image-relevant objects mentioned in the caption while UMIC shows limited sensitivity.
- All metrics are sensitive to the size of image-relevant objects mentioned in the caption.
- All metrics exhibit a weak understanding of negation.
- UMIC is sensitive to sentence structure, whereas CLIPScore and PAC-S demonstrate limited sensitivity.
- UMIC prioritizes correct sentence structure over mentions of larger objects or number of objection mentions in captions, whereas CLIPScore and PAC-S exhibit the opposite behavior.

Our primary contribution is highlighting specific areas where reference-free metrics exhibit limitations so that caution can be exercised when using these metrics for image captioning evaluation. We hope our findings will guide further improvements in reference-free evaluation of image captioning.

3.4. Related Works

Reference-free metrics: We study the robustness of CLIPScore [**37**], UMIC [**50**] and PAC-S [**79**]. CLIPScore measures the similarity between the image and the candidate caption using a scaled cosine similarity of the image and text representations from the CLIP [**71**] model. On the other hand, UMIC utilizes the UNITER [**19**] model, which is pre-trained to align image and text pairs, and finetunes it via contrastive learning to distinguish reference captions from its hard negatives. PAC-S [**79**] introduces a novel metric that strategically curates positive pairs for contrastive learning, enhancing the multimodal embedding space of CLIP. PAC-S employs scaled cosine similarity, akin to CLIPScore, to evaluate the similarity



Please summarize the question and answer in one sentence.
Question: What color is the table?
Answer: brown
Long answer: The color of table is brown.

Question: What color is the front of the train?
Answer: red and black
Long Answer: The color of the front of the train is red and black.

Question: What color of shirt is this tennis player wearing?
Answer: red

Long Answer: The color of the shirt this tennis player is wearing is red.

Fig. 3.2. Generating caption-like sentences by transforming visual question-answer pairs using GPT-J.

between the candidate caption and the provided image. SMURF [**30**] is another recently proposed metric for image caption evaluation, which has a reference-free evaluation of the fluency of the caption; however, the evaluation of the semantic correctness of the caption is still reference-based. Also, InfoMetIC [**38**] has the capability to pinpoint incorrect words and overlooked image areas at a fine-grained level while also providing an overall quality score at a coarse-grained level.

Vision-language benchmarks: Recently, a number of vision-language benchmarks have been proposed to evaluate the fine-grained understanding of relations, attributes, actions, and visio-linguistic compositionality in vision-language models, such as CAB [**103**], Winoground [**92**], ARO [**106**], VL-checklist [**111**], CREPE [**63**] and VALSE [**69**]. Although these evaluations also highlight the limitations of current models towards fine-grained understanding, our focus is specifically on evaluating the robustness of recently proposed reference-free image-captioning *metrics*. Our goal is to identify the scenarios where these metrics fail to distinguish between correct and incorrect captions to ensure the cautious use of these metrics in such scenarios.

3.5. Datasets Used to Conduct the Examination

3.5.1. Dataset Creation

To conduct our examination of the robustness of the metrics, we use a dataset of generated image captions. We generate image captions in one of the following ways, depending on the question we are trying to answer (see section 3.6 for more details):

QA to caption conversion: We employ GPT-J prompting to transform visual question-answer pairs into caption-like sentences. We used the questions from the popular VQAv2 [**32**] dataset. The answers could either be ground-truth answers or model generated depending on the analysis. Figure 3.2 shows an example caption-like sentence generated by GPT-J along with the prompt and support examples. The support examples are specific to the question type of the input question. More details about support example selection can be found in B.0.1).

Caption templates: We generate captions in a controlled setting in the format of the `"There is a/an [object name]."` for objects present in the image. We utilized the COCO detection dataset [**58**] to extract the names of objects in each image. This dataset provides object tags across 90 categories and attributes like objects' areas. The sentence construction process is elaborated within each baseline description.

We will make the dataset containing all the generated captions publicly available for the purpose of reproducibility and future use by the community.

3.5.2. Dataset Analysis

We conduct the following analyses of our generated captions dataset:

Human verification: We verify the quality of the generated captions, we conduct human verification. For a random subset of the captions generated via the *QA to caption conversion* method, we provide an expert human (graduate student) with the original question, the original answer and the generated caption. We ask the expert human to manually examine each caption and judge if the caption: 1) is grammatically correct, 2) contains all the information present in the original question-answer pair, 3) does not contain any hallucinated facts that are not present in the original question-answer pair. Upon examining 500 random samples, our analysis identified only 22 captions with grammatical errors, and only 14 captions with issues related to either hallucinating information or missing information from the original question and answer pairs.

We extended this analysis to 100 randomly sampled captions generated using the *caption template* method, and all samples were found to be correct, benefiting from their straightforward format.

Comparing generated captions with human written captions: For the captions generated using the *QA to caption conversion* method, it is worth asking how the distribution of such captions compares with that of human written captions in existing datasets, such as, COCO captions [**17**]. To throw light on this, we refer to [**6**] where they compared the distributions of nouns, verbs, and adjectives mentioned in COCO captions with those mentioned in the VQA questions and answers, and found that they are statistically significantly different from each other (Kolmogorov-Smirnov test, $p < 0.001$). Consequently, we

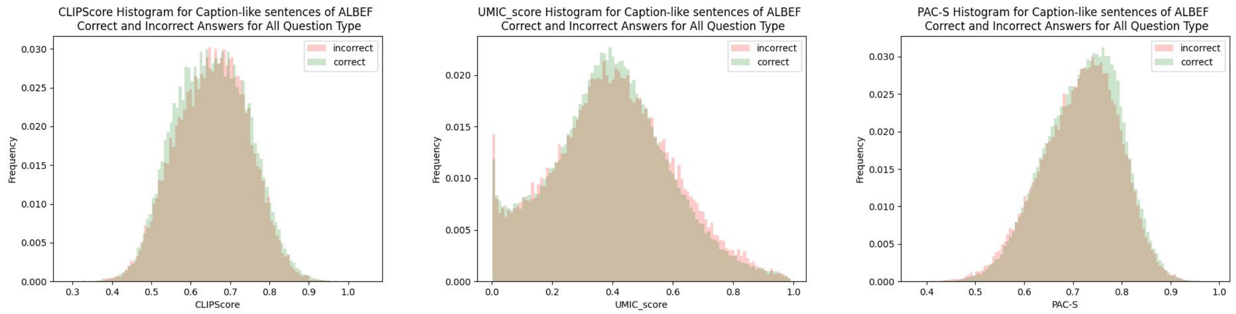


Fig. a: CLIPScore

Fig. b: UMIC

Fig. c: PAC-S

Fig. 3.3. Histograms of CLIPScore (Fig. a), UMIC (Fig. b), and PAC-S (Fig. c) for correct and incorrect caption-like sentences created using correct and incorrect answers from ALBEF for VQAv2 questions.

expect the captions generated through our *QA to caption conversion* method to exhibit different distributions of nouns, verbs, and adjectives compared to the human-written captions. However, [**6**] also show that the VQA questions and answers require a deeper understanding of images beyond what (human written) image captions typically capture. Thus, in spite of the differing word distributions between our generated captions and human written captions, we posit that our captions can be extremely valuable in **stress testing the robustness of image caption evaluation metrics**.

3.6. Experiments and Results

Preliminary experiment: First, we describe our preliminary experiment that served as a motivation for the rest of the study. We were interested in examining how different the scores assigned by reference-free image captioning metrics are for correct/incorrect captions created by converting questions and correct/incorrect answers from the VQAv2 dataset to caption-like sentences. Captions generated in this way are unique in that even for incorrect captions, a significant portion of it (corresponding to the question part) is still correct. Thus, such a dataset of captions serves as a good *stress test* dataset for examining the robustness of reference-free image captioning metrics.

To obtain correct and incorrect answers, we obtained predictions from the ALBEF [**53**] visual question answering model on the validation splits of the VQAv2[**32**] dataset. We fine-tuned ALBEF on this dataset and conducted IID evaluation. We then converted each question and its corresponding ALBEF answer into a caption-like sentence as described in Section 3.5. We only use answers that match with either three or more human answers (and we classify them as correct answers) or that do not match with any human answers (and we classify them as incorrect answers), resulting in a total of 179,297 answers (43389 incorrect and 135908 correct). The histograms of results for the VQAv2 dataset are presented in

Answer Type	**CLIPScore**	**UMIC**	**PAC-S**
VQAv2- Correct	0.480	0.394	0.558
VQAv2- Incorrect	0.481	0.403	0.549

Table 3.1. CLIPScore, UMIC, and PAC-S comparison for caption-like sentences for incorrect and correct answers generated by ALBEF model for VQAv2 dataset.

Answer Type	**CLIPScore**	**UMIC**	**PAC-S**
Correct yes/no	0.457	0.355	0.540
Incorrect yes/no	0.470	0.392	0.547
Correct numbers	0.468	0.354	**0.561**
Incorrect numbers	0.477	0.387	0.553
Correct others	0.512	0.452	0.578
Incorrect others	0.485	0.411	0.548

Table 3.2. CLIPScore, UMIC, and PAC-S comparison for correct and incorrect caption-like sentences generated with different answer types from VQAv2 dataset.

Figure 3.3. We see a significant overlap between the distributions of scores for correct and incorrect captions for all metrics, highlighting the limitations of these metrics in precisely assessing caption quality.

Score normalization: The UMIC final score, which is an output of a sigmoid function, has a value range between 0 and 1. On the other hand, the CLIPScore and PAC-S use the cosine similarity score scaled by a factor of 2.5 and 2, respectively. Although theoretically, CLIPScore can vary between -2.5 and 2.5, and PAC-S can vary between -2 and 2, we have not observed negative scores, and they rarely exceed 1.0. The distributions of metrics are illustrated in Figure 3.3. While we do not directly compare the values of these metrics in this paper, we aim to contrast their sensitivity to different factors. To achieve this, we apply the min-max normalization separately to each metric for every experiment. This method allows us to evaluate the respective sensitivities of the metrics effectively. Please note that all reported scores are normalized, but the histograms are plotted using the original scores to accurately represent the original distributions.

Score normalized results: As shown in Table 3.1, CLIPScore and UMIC assign higher average scores to incorrect captions compared to correct captions; however, PAC-S assigns higher average scores to correct captions. We conducted further analysis by examining the average scores assigned by these metrics for different answer types of the VQAv2 dataset (please refer to Table 3.2 for detailed scores). Specifically, we observed that for the 'yes/no' answer type, on average, all the metrics assign higher scores to incorrect captions. For the 'number' answer type, only PAC-S was able to assign higher average scores to correct captions. However, for the 'others' answer type, all the metrics assign higher average scores to correct captions.

Question Type	***CLIPScore Incorrect***	***CLIPScore Correct***	***UMIC Incorrect***	***UMIC Correct***	***PAC-S Incorrect***	***PAC-S Correct***
how many	0.475	0.468	0.372	0.354	0.559	0.562
what color	0.454	0.466	0.420	0.517	0.514	0.542
what sport	0.480	0.584	0.299	0.342	0.513	0.628
what animal	0.436	0.544	0.257	0.322	0.488	0.623
what time	0.469	0.405	0.333	0.282	0.528	0.492
what brand	0.440	0.458	0.481	0.511	0.497	0.508
what type/kind	0.485	0.537	0.382	0.417	0.544	0.594
where	0.501	0.551	0.380	0.435	0.561	0.620
which	0.495	0.529	0.419	0.414	0.556	0.581
what is/are the	0.497	0.543	0.436	0.468	0.559	0.605
others	0.480	0.471	0.412	0.370	0.549	0.550

Table 3.3. CLIPScore, UMIC, and PAC-S for correct and incorrect caption-like sentences generated for different question types of VQAv2.

For further investigation, we look at results for specific question types for VQAv2. As illustrated in Table 3.3), for CLIPScore, we observe that incorrect captions received higher scores on average for three question types: 'how many', 'what time' and 'others'. Also, UMIC assigns higher scores on average to incorrect captions for four question types: 'how many', 'what time', 'which', and 'others'. On the other hand, PAC-S assigns higher scores on average to incorrect captions for 'what time' and 'others' question types, suggesting **all metrics show poor performance for 'what time' questions**, which is considered to be a hard question type. Moreover, **CLIPScore and UMIC show poor performance for 'how many' questions.** Although PAC-S assigns higher average to correct captions over incorrect captions for 'how many' question type, the gap between the absolute values of average scores for correct and incorrect captions for 'how many' question is less than that for other question types.

Controlled investigation to identify sensitivity to various factors: Having established that these metrics struggle to distinguish the set of incorrect captions from the set of correct captions, in the following sections, we delve deeper into understanding the underlying reasons for their failure. To validate the comparisons made between different group means and ensure the reliability of our claims, we conducted a **t-test** for each comparison, using a p-value threshold of 0.01 (p-value < 0.01). Notably, all reported comparisons successfully satisfied this predetermined threshold, affirming the robustness of our statistical analyses.

3.6.1. Sensitivity to fine-grained errors

The primary objective of this section is to determine the sensitivity of these metrics to fine-grained errors. An incorrect caption is said to have "fine-grained errors" if it has high lexical overlap with a correct caption. To obtain such pairs of correct and incorrect captions,

Answer Type	CLIPScore	UMIC	PAC-S
Ground Truth	0.479	0.422	0.542
Incorrect from ALBEF	0.468	0.404	0.535

Table 3.4. CLIPScore, UMIC, and PAC-S comparison for caption-like sentences for incorrect answers generated by ALBEF model for VQAv2 and captions generated with its ground truth counterpart.

we first generate incorrect captions corresponding to the questions for which ALBEF produced incorrect responses. Then, we generate correct captions using ground-truth answers for the same set of questions. We convert the questions and answers into captions using the method described in Section 3.5. We excluded questions with yes/no answers from this study as we discuss them in Section 3.6.4. In total, we analyzed 38383 samples for this experiment.

We quantify the **degree of lexical overlap** between a pair of correct and incorrect captions in our dataset by measuring the F1 score between them. The mean F1 score across all such pairs in our dataset is 0.725. To place this in context, we measure the F1 score between pairs of correct (human-written) and incorrect (generated by image captioning models) captions from the Composite dataset [**1**], a widely-used dataset for evaluating image captioning metrics (see B.0.2 for more details on F1 score computation for Composite dataset). The mean F1 score across all such pairs from the Composite dataset is 0.224, which is significantly lower than that for our dataset. This highlights the difficulty of our dataset making it suitable for stress testing the robustness of image captioning metrics.

As demonstrated in Table 3.4, for all metrics, captions with ground truth answers received a higher average score compared to captions with fine-grained errors. Despite the higher average scores assigned to correct captions, the ranking results reveal that these metrics often fail to prioritize correct captions over incorrect ones. CLIPScore fails to rank correct captions above incorrect captions in 46.34% of cases, while UMIC fails to do so in 45.99% of cases. Also, PAC-S ranks incorrect captions over correct captions in 46.84% of times. Thus, **all metrics show weak sensitivity to detecting fine-grained errors**.

We also report a **human baseline** for the task of distinguishing correct captions from the ones with fine-grained errors. Presenting 1000 randomly sampled images, each with one correct and one incorrect caption, we instructed an expert human subject (graduate student) with the prompt: "An image is presented alongside two corresponding descriptions. Please identify the description that best aligns with the content depicted in the image.". The human subjects encountered difficulty ranking the correct caption above an incorrect one in only 16% of cases. Thus, the human performance is far better than that of automatic metrics .

Answer Type	**CLIPScore**	**UMIC**	**PAC-S**
Ground Truth	0.501	0.487	0.576
Plausible	0.474	0.242	0.527
Object from Image	0.526	0.354	0.601
Random	0.458	0.275	0.522

Table 3.5. CLIPScore, UMIC, and PAC-S comparison for caption-like sentences from VQAv2 ground truth, plausible, object from image and random answers.

3.6.2. Are metrics differently sensitive to different kinds of fine-grained errors?

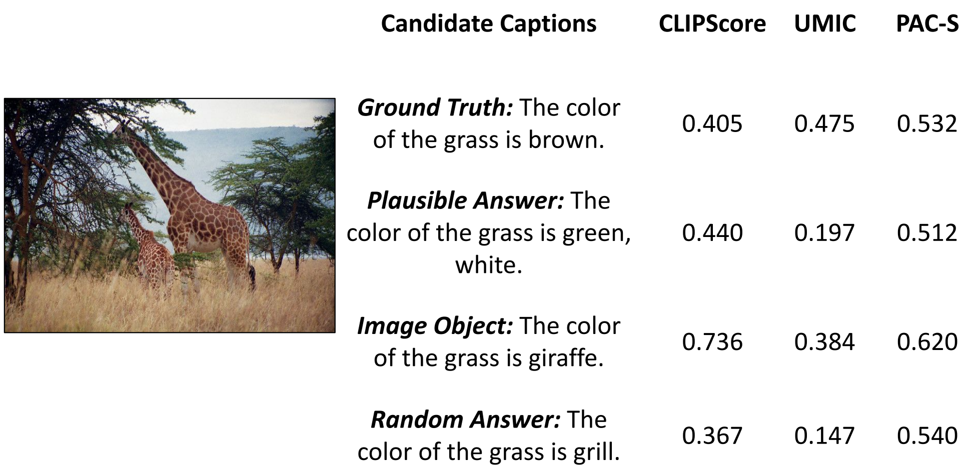
The main aim of this experiment is to assess if the metrics exhibit varying sensitivity to different types of fine-grained errors, in particular visual grounding errors and caption implausibility errors. To assess this, we generated three types of incorrect captions for each correct caption by replacing the ground-truth answer in the correct caption with: a plausible but incorrect answer (visual grounding error), an object found in the image (caption implausibility error), and a random answer (see Figure 3.4 for an example and see Appendix B.0.3 for more details on plausible answers).

For this experiment, we limited our investigation to the following question types: 'what number is', 'what time', 'what color', and 'what brand', as their answers are non-object entities and, therefore, are not present in the COCO Detection dataset. Thus, when constructing a sentence using an object in the image, we can be sure that it would result in an incorrect caption for the image. We analyzed 23841 sets of 4 captions each for this experiment.

As illustrated in Table 3.5, the score difference between the correct captions and the captions with implausibility errors is significantly smaller than the difference between the correct captions and the captions with visual grounding errors. This indicates that the metrics exhibit **lower sensitivity** to caption **implausibility errors** and **higher sensitivity** to **visual grounding errors**. Notably, both CLIPScore and PAC-S assigned higher average scores to captions with implausibility errors compared to ground truth answers, and only UMIC assigned higher average score to captions with ground truth answers. In the following sections, we further examine the sensitivity of the metrics to various visual and linguistic aspects.

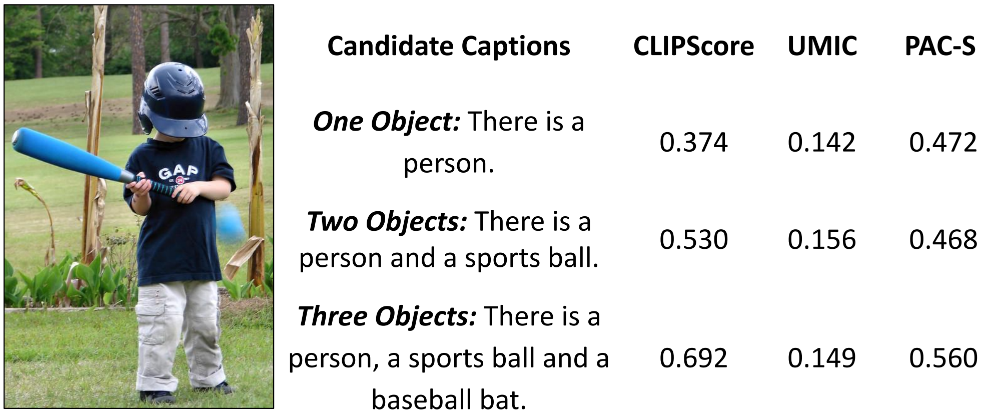
3.6.3. Visual Aspects

3.6.3.1. **Sensitivity to the number of object mentions in the caption.** In this section, our objective is to assess the sensitivity of the metrics to the number of objects mentioned in the caption. To conduct this evaluation, we filter images from COCO Detection dataset [**58**] having a minimum of three object tags and randomly select three object tags for



Candidate Captions	CLIPScore	UMIC	PAC-S
Ground Truth: The color of the grass is brown.	0.405	0.475	0.532
Plausible Answer: The color of the grass is green, white.	0.440	0.197	0.512
Image Object: The color of the grass is giraffe.	0.736	0.384	0.620
Random Answer: The color of the grass is grill.	0.367	0.147	0.540

Fig. 3.4. Captions from ground truth, plausible answer, an object from the image and a random asnwer of VQAv2.



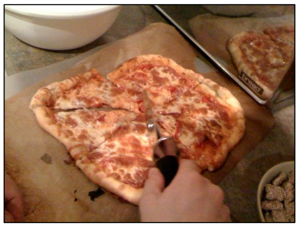
Candidate Captions	CLIPScore	UMIC	PAC-S
One Object: There is a person.	0.374	0.142	0.472
Two Objects: There is a person and a sports ball.	0.530	0.156	0.468
Three Objects: There is a person, a sports ball and a baseball bat.	0.692	0.149	0.560

Fig. 3.5. Captions referring to different number of objects from the image.

Number of Objects	CLIPScore	UMIC	PAC-S
One Object	0.449	0.205	0.500
Two Objects	0.512	0.212	0.540
Three Objects	0.561	0.195	0.578
Shuffled One Object	0.445	0.139	0.503
Shuffled Two Objects	0.499	0.148	0.541
Shuffled Three Objects	0.540	0.169	0.576

Table 3.6. CLIPScore, UMIC, and PAC-S comparison for sentences with various number of objects name, and their shuffled counterparts.

each image and utilize their corresponding object names to form sentences, depicting one, two, and three objects presented in the image (see Figure 3.5). We analyzed 19412 images for this experiment.



Candidate Captions	CLIPSore	UMIC	PAC-S
Small Object: There is a knife.	0.460	0.507	0.561
Big Object: There is a pizza.	0.632	0.469	0.718
Shuffled Small Object: A there knife is.	0.480	0.268	0.561
Shuffled Big Object: A there pizza is.	0.664	0.250	0.719

Fig. 3.6. Captions referring to small and large area of the image and their shuffled counterparts.

Object Size	CLIPScore	UMIC	PAC-S
Small Object	0.396	0.317	0.492
Big Object	0.434	0.232	0.580
Shuffled Small Object	0.390	0.205	0.495
Shuffled Big Object	0.436	0.170	0.590

Table 3.7. CLIPScore, UMIC, and PAC-S comparison for captions referring to small and a big objects in the image, and their shuffled counterparts.

As presented in the first three rows of Table 3.6, CLIPScore and PAC-S scores for captions with three objects are significantly higher than for captions with two objects. Also, captions with two objects score significantly higher than those with one object. In contrast, for UMIC, captions with one, two, and three objects received average scores of 0.205, 0.212, and 0.195, respectively. Although the t-test indicated statistically significant differences between scores across different object counts, the gap between absolute score values is smaller for UMIC than for CLIPScore and PAC-S. In conclusion, **CLIPScore and PAC-S display a heightened sensitivity to the number of image-relevant objects mentioned in the caption, while UMIC shows limited sensitivity towards this factor**.

3.6.3.2. **Sensitivity to size of objects mentioned in the caption**. In this experiment, our primary goal is to examine the effect of object size mentioned in captions on the metrics. To achieve this, we utilize the COCO Detection dataset [**58**] to select one small and one large object from the same image with a noticeable difference in the area (see Figure 3.6 for an example and for detailed explanation see Appendix B.0.4.). As a result, we selected 24610 images for further analysis.

As demonstrated in the first two rows of Table, 3.7, for CLIPScore and PAC-S, captions with smaller objects received a lower average score than those with bigger objects. On the other hand, UMIC assigned a higher average score to captions with smaller objects compared to captions with bigger objects. Overall, **all metrics demonstrate sensitivity to the size of image-relevant objects mentioned in the caption**.

3.6.4. Linguistic Aspects

3.6.4.1. **Sensitivity to negation**. To assess the ability of metrics to distinguish between correct captions and their negated versions, we created 80530 captions-like sentences by using the questions with 'yes' or 'no' ground-truth answers from the validation split of VQAv2. Additionally, we generated negated captions by negating the ground truth answer.

For CLIPScore, correct captions received a higher score of 0.457, and their negated versions got 0.450 on average. For UMIC, correct captions received a higher average of 0.359, and their negated versions got 0.335 on average. Correct captions received a higher average of 0.556 for PAC-S, and their negated versions got 0.548 on average. Although the correct captions scored statistically significantly higher than the negated ones, CLIPScore, UMIC, and PAC-S ranked the negated caption above the correct caption incorrectly in 41.36%, 44.24%, and 41.83% of cases, respectively. Thus, **all metrics exhibit a weak understanding of negation**.

3.6.4.2. **Sensitivity to the sentence structure**. To evaluate the sensitivity of the metrics to sentence structure, we generated 214354 caption-like sentences with VQAv2 ground truth answers and then shuffled them. For CLIPScore, correct captions received 0.469, and their shuffled version got 0.450 on average. For UMIC, correct captions received 0.400, and their shuffled version got 0.211 on average. Correct captions received 0.548 for PAC-S, and their shuffled version got 0.539 on average. Despite higher average scores assigned to correct captions, the ranking results reveal that CLIPScore fails to rank the correct caption higher than the shuffled one in 34.32% of cases, contrasting with UMIC, where this occurs in only 9.18% of cases. Additionally, PAC-S falls short, assigning a higher score to the correct caption than the shuffled one in 43.05% of cases. This indicates that **UMIC is more responsive to the structure of the sentence compared to CLIPScore and PAC-S**.

3.6.5. Visio-Linguistic Aspects

3.6.5.1. **Sentence Structure versus Visual Aspects**. To evaluate the sensitivity of the metrics to sentence structure, we generated 214354 caption-like sentences with VQAv2 ground truth answers and then shuffled them. For CLIPScore, correct captions received 0.469, and their shuffled version got 0.450 on average. For UMIC, correct captions received 0.400, and their shuffled version got 0.211 on average. Correct captions received 0.548

for PAC-S, and their shuffled version got 0.539 on average. Despite higher average scores assigned to correct captions, the ranking results reveal that CLIPScore fails to rank the correct caption higher than the shuffled one in 34.32% of cases, contrasting with UMIC, where this occurs in only 9.18% of cases. Additionally, PAC-S falls short, assigning a higher score to the correct caption than the shuffled one in 43.05% of cases. This indicates that **UMIC is more responsive to the structure of the sentence compared to CLIPScore and PAC-S**.

3.7. Conclusion and Discussion

In conclusion, recently proposed reference-free image captioning evaluation metrics are far from perfect; they cannot distinguish an incorrect caption from a correct caption when the difference between them is fine-grained. The sensitivity of CLIPScore, UMIC, and PAC-S varies across different error types: they are less affected by plausibility errors yet more by visual grounding errors. All metrics struggle with understanding negation. All metrics are influenced by the size of the relevant objects mentioned in the caption, and CLIPScore and PAC-S also responds to the number of object mentions. UMIC is responsive to sentence structure, while CLIPScore and PAC-S disregards it often. Moreover, UMIC prioritizes sentence structure over the number and size of objects mentioned in the caption; in contrast CLIPScore and PAC-S prioritize the object size and number of object mentions over sentence structure.

Our primary contribution is highlighting specific areas where reference-free metrics exhibit limitations. The root cause of these limitations is traced to the insufficient fine-grained understanding of the CLIP and UNITER models upon which these reference-free metrics rely. Promising avenues for enhancing this understanding include exploring object-centric representations and incorporating training with hard negatives [**106, 107, 12**]. Given the restricted fine-grained understanding of the underlying models shaping these metrics, caution is advised when employing them as evaluation metrics for image captioning.

3.8. Limitations

As a limitation, it is important to consider that responses marked as incorrect may not always be incorrect due to the stringent nature of VQA evaluation metrics [**3**]. Our approach does not account for this factor. However, for our experiments, since we fine-tune ALBEF for each domain, the risk of this issue is low. To get a quantitative sense, we randomly sampled 100 incorrect answers (as deemed by the VQA automatic metric) generated by ALBEF for VQAv2, and in only 10% of cases, the answer was actually correct (as deemed by an expert human). Furthermore, it is important to note that we do not account for the saliency of objects mentioned in the caption, which could be a confounding factor in our evaluation.

3.9. Ethics Statement

To enhance transparency and explainability, we conducted experiments aimed at shedding light on the evaluation process of the metric. By doing so, we aimed to provide insights and explanations that enable users to better comprehend and trust the metric's evaluations. Furthermore, we evaluated the robustness of the metrics, contributing towards the development of less biased evaluation metrics.

While we assess various aspects of existing metrics, it is important to note that our evaluation does not specifically examine metrics' potential biases across different demographics, including gender or race. While our research does not include an explicit experiment on bias perpetuation or amplification, we strongly encourage future studies to investigate how metrics may interact with biases present in datasets. This research direction is crucial in developing metrics that are less biased and more inclusive towards diverse demographics.

Chapter 4

Conclusion

This thesis focuses on delving into a multimodal Artificial Intelligence (AI) task known as Image Captioning. Image Captioning represents a crucial intersection of computer vision and natural language processing, aiming to bridge the semantic understanding gap between visual content and textual representation. The complexity of this task lies in the ability to not only recognize and interpret diverse visual elements within an image but also to convey this understanding in a linguistically coherent and contextually relevant manner.

The primary contribution of this thesis, MAPL [**65**], introduces a parameter-efficient method for repurposing pre-trained unimodal models for multimodal tasks including image captioning. The method, called MAPL, learns a lightweight mapping between the representation spaces of a pre-trained vision encoder and a pre-trained language model using aligned image-text data. This approach allows the model to generalize to unseen vision-language tasks with only a few in-context examples, making it effective for low-data and in-domain learning.

There are several promising avenues for future research in modeling vision and language, which we will highlight. One potential avenue for future research into pre-training vision and language models is the learning of fine-grained representations. Including fine-grained details is critical for enriching semantic information, improving generalization across diverse datasets, increasing robustness to variations in input data, and ensuring better alignment between visual and linguistic modalities. Exploring methods to integrate fine-grained features during pre-training effectively will contribute to advancing vision and language models, enabling them to achieve more nuanced understanding and improved performance in various applications, including image captioning. One other promising research avenue for Vision and Language Models (VLMs) is to learn efficiently with limited data. Present methodologies often entail training VLMs with substantial data and resource-intensive computations, posing sustainability challenges. A practical resolution involves the development of effective VLMs using constrained image-text data. Instead of relying solely on individual image-text pairs, a more valuable approach is integrating supervision across multiple image-text pairs, providing richer insights. One potential area for further investigation is pre-training Vision and Language Models (VLMs) with multiple languages. This strategy aims to address biases inherent in the prevalent practice of training these models with a single language, often English. By doing so, the models can acquire a broader understanding of diverse cultural visual characteristics linked to the same word meanings across different languages.

Our second study investigates the robustness of reference-free evaluation metrics for image captioning. Traditionally, the evaluation of image caption quality has been based on a reference-based methodology that measures the lexical overlap between generated and reference captions. However, this approach has limitations, as the set of references may not encompass the full range of acceptable captions. Additionally, metrics based on lexical overlap have a tendency to favor captions with similar vocabulary but very different meanings. To overcome these constraints, recent metrics, exemplified by CLIPScore [**37**], UMIC [**50**], and PAC-S [**79**], have introduced reference-free methodologies for evaluating image caption quality, aligning more closely with human judgments. Our primary contribution is highlighting specific areas where reference-free metrics exhibit limitations. Despite their strong correlation with human judgments, our findings indicate that CLIPScore, UMIC, and PAC-S encounter challenges in pinpointing fine-grained errors. These constraints stem from the insufficient fine-grained understanding of the CLIP and UNITER models, upon which the foundation of these reference-free metrics relies. Promising avenues for enhancing this understanding include exploring object-centric representations and incorporating training with hard negatives [**106, 107, 12**]. Given the restricted fine-grained understanding of the underlying models shaping these metrics, caution is advised when employing them as evaluation metrics for image captioning.

We anticipate that our insights will pave the way for further refinements in the parameter-efficient training and reference-free evaluation of image captioning.

References

[1] Somak ADITYA, Yezhou YANG, Chitta BARAL, Yiannis ALOIMONOS et Cornelia FERMÜLLER : Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173, 12 2017.

[2] Aishwarya AGRAWAL, Ivana KAJIĆ, Emanuele BUGLIARELLO, Elnaz DAVOODI, Anita GERGELY, Phil BLUNSOM et Aida NEMATZADEH : Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization. *arXiv preprint arXiv:2205.12191*, 2022.

[3] Aishwarya AGRAWAL, Ivana KAJIC, Emanuele BUGLIARELLO, Elnaz DAVOODI, Anita GERGELY, Phil BLUNSOM et Aida NEMATZADEH : Reassessing evaluation practices in visual question answering: A case study on out-of-distribution generalization. *In Findings of the Association for Computational Linguistics: EACL 2023*, pages 1201–1226, Dubrovnik, Croatia, mai 2023. Association for Computational Linguistics.

[4] Jean-Baptiste ALAYRAC, Jeff DONAHUE, Pauline LUC, Antoine MIECH, Iain BARR, Yana HASSON, Karel LENC, Arthur MENSCH, Katherine MILLICAN, Malcolm REYNOLDS, Roman RING, Eliza RUTHERFORD, Serkan CABI, Tengda HAN, Zhitao GONG, Sina SAMANGOOEI, Marianne MONTEIRO, Jacob MENICK, Sebastian BORGEAUD, Andrew BROCK, Aida NEMATZADEH, Sahand SHARIFZADEH, Mikolaj BINKOWSKI, Ricardo BARREIRA, Oriol VINYALS, Andrew ZISSERMAN et Karen SIMONYAN : Flamingo: a visual language model for few-shot learning. *In* Alice H. OH, Alekh AGARWAL, Danielle BELGRAVE et Kyunghyun CHO, éditeurs : *Advances in Neural Information Processing Systems*, 2022.

[5] Jean-Baptiste ALAYRAC, Jeff DONAHUE, Pauline LUC, Antoine MIECH, Iain BARR, Yana HASSON, Karel LENC, Arthur MENSCH, Katie MILLICAN, Malcolm REYNOLDS *et al.* : Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[6] Stanislaw ANTOL, Aishwarya AGRAWAL, Jiasen LU, Margaret MITCHELL, Dhruv BATRA, C. ZITNICK et Devi PARIKH : Vqa: Visual question answering. 2, 05 2015.

[7] Stanislaw ANTOL, Aishwarya AGRAWAL, Jiasen LU, Margaret MITCHELL, Dhruv BATRA, C Lawrence ZITNICK et Devi PARIKH : Vqa: Visual question answering. *In Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[8] Satanjeev BANERJEE et 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, Ann Arbor, Michigan, juin 2005. Association for Computational Linguistics.

[9] Hangbo BAO, Li DONG, Songhao PIAO et Furu WEI : BEit: BERT pre-training of image transformers. *In International Conference on Learning Representations*, 2022.

[10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill *et al.* : On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[11] 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.

[12] Emanuele Bugliarello, Aida Nematzadeh et Lisa Anne Hendricks : Weakly-supervised learning of visual relations in multimodal pretraining, 2023.

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov et Sergey Zagoruyko : End-to-end object detection with transformers. *In European conference on computer vision*, pages 213–229. Springer, 2020.

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski et Armand Joulin : Emerging properties in self-supervised vision transformers. *In Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[15] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu et Bo Xu : Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

[16] Jun Chen, Han Guo, Kai Yi, Boyang Li et Mohamed Elhoseiny : Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. *arXiv preprint arXiv:2102.10407*, 2021.

[17] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár et C Lawrence Zitnick : Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[18] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng et Jingjing Liu : Uniter: Learning universal image-text representations. 2019.

[19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng et Jingjing Liu : Uniter: Universal image-text representation learning. *In European conference on computer vision*, pages 104–120. Springer, 2020.

[20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng et Jingjing Liu : *UNITER: UNiversal Image-TExt Representation Learning*, pages 104–120. 09 2020.

[21] Jaemin Cho, Jie Lei, Hao Tan et Mohit Bansal : Unifying vision-and-language tasks via text generation. *In International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[22] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma *et al.* : Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[23] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu et Pascale Fung : Enabling multimodal generation on CLIP via vision-language knowledge distillation. *In Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland, mai 2022. Association for Computational Linguistics.

[24] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh et Dhruv Batra : Visual dialog. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova : BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, juin 2019. Association for Computational Linguistics.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly *et al.* : An image is worth 16x16 words: Transformers for image recognition at scale. *In ICLR*, 2020.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit et Neil Houlsby : An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations*, 2021.

[28] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu et Michael Zeng : An empirical study of training end-to-end vision-and-language transformers. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, June 2022.

[29] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu et Anette Frank : Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021.

[30] Joshua Feinglass et Yezhou Yang : Smurf: Semantic and linguistic understanding fusion for caption evaluation via typicality analysis. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2021.

[31] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima *et al.* : The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[32] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra et Devi Parikh : Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398–414, 2016.

[33] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo et 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.

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

[35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar et Ross B. Girshick : Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun : Deep residual learning for image recognition. pages 770–778, 06 2016.

[37] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras et Yejin Choi : CLIPScore: A reference-free evaluation metric for image captioning. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, novembre 2021. Association for Computational Linguistics.

[38] Anwen Hu, Shizhe Chen, Liang Zhang et Qin Jin : Infometic: An informative metric for reference-free image caption evaluation, 2023.

[39] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu et Jianlong Fu : Seeing out of the box: End-to-end pre-training for vision-language representation learning. *In Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12976–12985, June 2021.

[40] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu et Jianlong Fu : Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020.

[41] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li et 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.

[42] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li et Tom Duerig : Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[43] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen et Xiang Ren : A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland, mai 2022. Association for Computational Linguistics.

[44] Andrej Karpathy et Li Fei-Fei : Deep visual-semantic alignments for generating image descriptions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[45] Andrej Karpathy et Fei-Fei Li : Deep visual-semantic alignments for generating image descriptions. *In CVPR*, pages 3128–3137. IEEE Computer Society, 2015.

[46] Wonjae Kim, Bokyung Son et Ildoo Kim : Vilt: Vision-and-language transformer without convolution or region supervision. *In International Conference on Machine Learning*, 2021.

[47] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, Michael Bernstein et Fei-Fei Li : Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123, 05 2017.

[48] Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton : Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[49] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville et Yoshua Bengio : Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

[50] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui et Kyomin Jung : UMIC: An unreferenced metric for image captioning via contrastive learning. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226, Online, août 2021. Association for Computational Linguistics.

[51] Junnan Li, Dongxu Li, Caiming Xiong et Steven Hoi : Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *In ICML*, 2022.

[52] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong et Steven Chu Hong Hoi : Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[53] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong et Steven Hoi : Align before fuse: Vision and language representation learning with momentum distillation. *In NeurIPS*, 2021.

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

[55] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi et Jianfeng Gao : Oscar: Object-semantics aligned pre-training for vision-language tasks. *In ECCV*, August 2020.

[56] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei *et al.* : Oscar: Object-semantics aligned pre-training for vision-language tasks. *In European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[57] Chin-Yew Lin : ROUGE: A package for automatic evaluation of summaries. *In Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, juillet 2004. Association for Computational Linguistics.

[58] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C. Lawrence Zitnick : Microsoft coco: Common objects in context. *In European Conference on Computer Vision*, 2014.

[59] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin et Baining Guo : Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[60] Jiasen Lu, Dhruv Batra, Devi Parikh et Stefan Lee : *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

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

[62] Ziyang Luo, Yadong Xi, Rongsheng Zhang et Jing Ma : VC-GPT: visual conditioned GPT for end-to-end generative vision-and-language pre-training. *CoRR*, abs/2201.12723, 2022.

[63] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao et Ranjay Krishna : Crepe: Can vision-language foundation models reason compositionally?, 2023.

[64] Kenneth Marino, Mohammad Rastegari, Ali Farhadi et 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.

[65] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal et Aishwarya Agrawal : Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting, 2023.

[66] Jack Merullo, Louis Castricato, Carsten Eickhoff et Ellie Pavlick : Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

[67] Ron Mokady, Amir Hertz et Amit H Bermano : Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[68] Kishore Papineni, Salim Roukos, Todd Ward et 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, Philadelphia, Pennsylvania, USA, juillet 2002. Association for Computational Linguistics.

[69] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto et Albert Gatt : VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. *In Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 8253–8280, Dublin, Ireland, mai 2022. Association for Computational Linguistics.

[70] 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.

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger et Ilya Sutskever : Learning transferable visual models from natural language supervision. *In International Conference on Machine Learning*, 2021.

[72] Alec Radford, Karthik Narasimhan, Tim Salimans et Ilya Sutskever : Improving language understanding by generative pre-training. 2018.

[73] 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.

[74] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase et Yuxiong He : Zero: Memory optimizations toward training trillion parameter models. *In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[75] Shaoqing Ren, Kaiming He, Ross Girshick et Jian Sun : Faster R-CNN: Towards real-time object detection with region proposal networks. *In Advances in Neural Information Processing Systems (NIPS)*, 2015.

[76] Laria Reynolds et Kyle McDonell : Prompt programming for large language models: Beyond the few-shot paradigm. *In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[77] David Rolnick, Andreas Veit, Serge Belongie et Nir Shavit : Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[78] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg et Li Fei-Fei : Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 09 2014.

[79] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi et Rita Cucchiara : Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[80] Timo Schick et Hinrich Schütze : It's not just size that matters: Small language models are also few-shot learners. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, 2021.

[81] Piyush Sharma, Nan Ding, Sebastian Goodman et 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.

[82] Peter Shaw, Jakob Uszkoreit et Ashish Vaswani : Self-attention with relative position representations. *In NAACL-HLT (2)*, 2018.

[83] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao et Kurt Keutzer : How much can CLIP benefit vision-and-language tasks? *In International Conference on Learning Representations*, 2022.

[84] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach et Amanpreet Singh : Textcaps: a dataset for image captioning with reading comprehension. *In European Conference on Computer Vision*, pages 742–758. Springer, 2020.

[85] Karen Simonyan et Andrew Zisserman : Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[86] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh et Marcus Rohrbach : Towards vqa models that can read. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[87] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei et Jifeng Dai : Vl-bert: Pre-training of generic visual-linguistic representations. *In International Conference on Learning Representations*, 2020.

[88] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai et Yoav Artzi : A corpus for reasoning about natural language grounded in photographs. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

[89] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke et Andrew Rabinovich : Going deeper with convolutions. pages 1–9, 06 2015.

[90] Hao Tan et Mohit Bansal : LXMERT: Learning cross-modality encoder representations from transformers. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, novembre 2019. Association for Computational Linguistics.

[91] Hao Tan et Mohit Bansal : Lxmert: Learning cross-modality encoder representations from transformers. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[92] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela et Candace Ross : Winoground: Probing vision and language models for visio-linguistic compositionality. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022.

[93] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles et Herve Jegou : Training data-efficient image transformers &amp; distillation through attention. *In* Marina Meila et Tong Zhang, éditeurs : *Proceedings of the 38th International Conference on Machine Learning*, volume 139 de *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

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

[95] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals et Felix Hill : Multimodal few-shot learning with frozen language models. *In* M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang et J. Wortman Vaughan, éditeurs : *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc., 2021.

[96] Vladimir N Vapnik et A Ya Chervonenkis : On the uniform convergence of relative frequencies of events to their probabilities. *In Measures of complexity*, pages 11–30. Springer, 2015.

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser et Illia Polosukhin : Attention is all you need. *In Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[98] Ramakrishna Vedantam, C. Lawrence Zitnick et Devi Parikh : Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2014.

[99] Ben Wang et Aran Komatsuzaki : GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, mai 2021.

[100] Ben Wang et Aran Komatsuzaki : GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, mai 2021.

[101] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov et Yuan Cao : Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021.

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

[103] Yutaro Yamada, Yingtian Tang et Ilker Yildirim : When are lemons purple? the concept association bias of clip, 2022.

[104] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu et Lijuan Wang : An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*, 2021.

[105] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu et Lijuan Wang : An empirical study of gpt-3 for few-shot knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:3081–3089, 06 2022.

[106] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky et James Zou : When and why vision-language models behave like bags-of-words, and what to do about it? *In The Eleventh International Conference on Learning Representations*, 2023.

[107] Le Zhang, Rabiul Awal et Aishwarya Agrawal : Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding, 2023.

[108] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi et Jianfeng Gao : Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584, 2021.

[109] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi et Jianfeng Gao : Vinvl: Revisiting visual representations in vision-language models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[110] 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.

[111] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu et Jianwei Yin : An explainable toolbox for evaluating pre-trained vision-language models. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE, décembre 2022. Association for Computational Linguistics.

[112] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso et Jianfeng Gao : Unified vision-language pre-training for image captioning and vqa. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

Appendix A

Appendix for MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting

A.0.1. Leveraging more shots

In Table 2.2, we observe MAPL's performance rapidly plateaus as the number of few-shot examples increases beyond 4. We hypothesize this could be related to the mapping network being trained on single image-caption pairs, and/or the visual embeddings still not being fully in-distribution with the language embeddings. Intuitively, a handful of examples may often help with task location [**76**]; however, the more shots are added, the more out-of-distribution the multimodal prompt becomes. This issue could be mitigated with in-context example selection [**104**] or better mixing of visual and textual modalities.

A.0.2. 0-shot VQAv2 results with OPT

In Table 2.5, we observe Frozen*-OPT outperforms MAPL-OPT on 0-shot VQAv2. Upon close inspection, we notice MAPL-OPT often generates longer answers for yes/no questions, which receive a score of 0 according to VQA accuracy and VQAv2 reference answers – this is a problem of the metric and not the model itself [**2**]. After filtering all answers starting with "yes" or "no" to leave only the short answer, MAPL-OPT achieves a VQA accuracy of 40.14% while Frozen*-OPT only reaches 32.03%.

A.0.3. 4-shot results with OPT

In Table 2.5, we also observe that few-shot VQA performance is considerably lower for configurations using OPT-6.7B as language model. This is possibly due to the lack of a relative positional encoding [**82**] in OPT, which is required for the transformer to generalize

	Ablated	**Original**	**Changed**	**VQAv2**	**OK-VQA**	**TextVQA**	**VizWiz-VQA**	**CC**	**COCO**	**TextCaps**	**VizWiz-Caps**	**Overall Δ**	
	setting	**value**	**value**	**4-shot**	**4-shot**	**4-shot**	**4-shot**	**CIDEr**	**CIDEr**	**CIDEr**	**CIDEr**	**4-shot**	**CIDEr**
		MAPL		46.39	25.49	9.87	20.02	71.90	54.90	23.30	21.70	0	0
(i)	Vision encoder	CLIP-ViT-L/14	CLIP-ViT-B/32	43.48	24.43	7.85	15.97	59.20	47.00	19.00	15.70	-2.51	-7.73
(ii)	Visual features	Grid	Global	43.75	22.90	8.81	18.20	66.70	49.70	18.40	19.90	-2.03	-4.28
(iii)	Mapping network size	Medium	Small	44.83	26.37	9.68	17.78	68.30	55.50	21.30	20.80	-0.78	-1.48
			Large	45.03	23.92	8.88	19.01	73.40	57.10	24.00	23.10	-1.23	+1.45
(iv)	Output seq. length	32	16	44.18	25.16	9.01	18.15	72.80	56.20	22.50	21.80	-1.32	+0.37
			64	45.22	25.07	10.35	18.89	74.80	58.30	24.30	24.80	-0.56	+2.60
(v)	Learned constant embeddings	Yes	No	40.87	19.31	11.42	16.79	80.52	57.49	30.07	26.16	-3.35	+5.61
(vi)	Data quality Visual features	Clean Grid	Noisy Global	42.80	22.59	6.06	17.33	93.80	42.10	15.60	15.70	-3.25	-1.15

Table A.1. Ablation studies. "Overall Δ" refers to the difference (ablated model - base model), averaged across datasets per task.

to prompt sequences where an image is not always in the first absolute position, or which contain more than one image [**94**].

A.0.4. Implementation details on simpler mapping networks

In Sec. 2.6.5, we ablate the choice of mapping network architecture and replace it by simpler architectures. Similarly to [**29, 66**], the linear mapping is applied per-position on top of a flattened grid of visual features, and it projects from $D_i = 1024$ to $D_o = 4096$ dimensions (4.2M parameters). The output sequence length L_o is thus equal to $L_i = 257$ (instead of 32) – as explained in Sec. 2.5.1, this increases the computational complexity in the subsequent LM, which in turn increases training and inference time considerably. Similarly to [**67**], the 2-layer MLP is applied on top of a global vector of visual features. The MLP's hidden dimensionality D_h is equal to $D_i = 1024$, and the output dimensionality is $D_o = 32 * 4096$, which we split into 32 vectors of 4096 dimensions (135.3M parameters).

A.0.5. Additional ablation studies

Table A.1 shows the results of our additional ablation studies. Unless specified otherwise, we perform all ablations on $\text{MAPL}_{\text{CC-clean}}$ trained with 100% of the data. These experiments are run only once and early stopping is based on the validation split of Conceptual Captions.
Pre-trained vision encoder. We ablate the pre-trained vision encoder used to compute image representations. We report results in row (i) of Table A.1. We compare two CLIP [**70**] variants, our choice based on the ViT-L/14 backbone and the ViT-B/32 backbone. Indeed, the ViT-L/14 based vision encoder has an average +20% advantage over the ViT-B/32 variant. We hypothesize this improvement is probably due to finer-grained image patches and a bigger model size.

Global vs. grid visual features. Grid features – as opposed to global features – preserve the spatial information in images. This kind of fine-grained information might be useful for complex VL tasks. To measure the impact of grid features, we train a version of MAPL where we use the global image representation from CLIP's multimodal embedding space. Results are reported in row (ii) of Table A.1. We observe an average -10% drop in performance, validating our choice of using grid over global visual features.

Mapping network architecture. We ablate the architectural design of our mapping network in rows (iii) and (v) of Table A.1. First, we ablate the size of our mapping network in terms of depth and hidden size. We explore three options: Small (2 layers and hidden size of 128), Medium (4 layers and hidden size of 256), and Large (8 layers and hidden size of 512). We see that using a smaller mapping network generally performs slightly worse than the base model. On the other hand, using a larger mapping network improves only in image captioning tasks, while increasing significantly the number of trainable parameters (from 3.4M to 19.5M). We also ablate the output sequence length L_o of our mapping network. Similarly, reducing the output sequence length to 16 yields slightly lower performance overall, and increasing it to 64 only improves in image captioning tasks. In the extreme, we completely remove the learned constant embeddings and output the same sequence length coming from the vision encoder, i.e., $L_o = L_i = 257$. Following the trend, increasing the number of mapped visual embeddings is beneficial for image captioning but hurts VQA performance, while notably reducing training and inference throughput.

Data quality & visual features. This ablation setting aims to be the most similar to Frozen: training on the *full* (noisy) Conceptual Captions dataset while using global visual features. Results are reported in row (vi) of Table A.1. The overall performance is worse than that of our base model (-16% on average), but still better than Frozen* $_{\text{CC}}$ on Table 2.4 (+81% on average). This validates our choice of using grid visual features while training on a subset of cleaner data.

A.0.6. Additional qualitative results

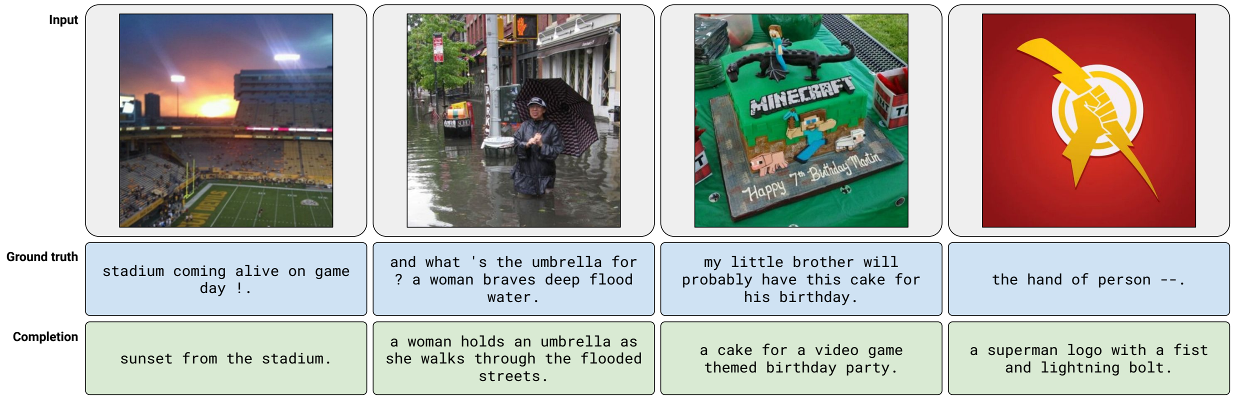
Figures A.1-A.12 show additional qualitative results of MAPL$_{\text{CC-clean}}$ on random samples from different image captioning and VQA datasets. For VQA, in-context learning from 4 shots is performed.

A.0.7. Interpretability of visual embeddings

Using MAPL$_{CC-clean}$, we extract mapped visual embeddings (after the mapping network) for ∼30 images from the COCO Karpathy-test set, and compute the nearest token embeddings (from the LM's vocabulary) using cosine similarity. We rarely found the top-5 nearest tokens correspond to concepts present in the image, suggesting these embeddings are not interpretable. We hypothesize this is perhaps because they carry a combination of task-inducing and image-specific information, also pointed out by [**67**]. We further cluster the mapped visual embeddings with K-means, and observe that each cluster often represents some visual concept (e.g., animals, food, sports). This means the mapped visual embeddings retain visual information from the vision encoder, which we also verify with MAPL's performance on VL benchmarks.

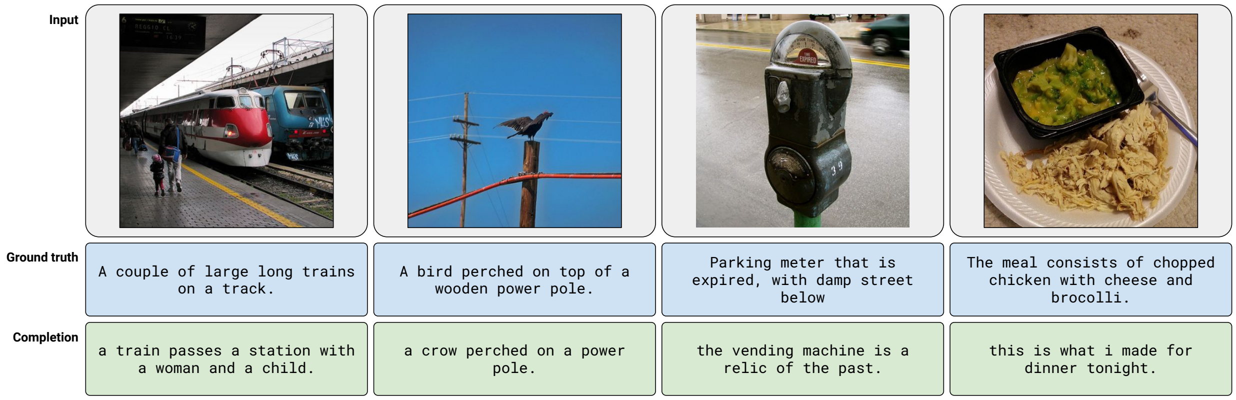
A.0.8. Analysis of VQA answer distributions

In this section, we show the distribution of answers for selected VQAv2 question types. We compare MAPL with several baselines of our model to get insights into how the model's predictions change when training on increasing multimodal data. For the text-only baseline, we only provide the question text to the LM. This is different from the previously introduced blind baseline (Sec. 2.6.1), where a blacked-out image is also provided. In particular, we compare the predicted answer distribution of $\text{MAPL}_{\text{COCO}}$ evaluated on on zero- and few-shot VQA with the aforementioned baselines and the ground truth. Overall, we observe the predicted answer distribution gets closer to the ground truth answer distribution (Figure A.18) as more information from the image-question pair is provided to the model. We notice a considerable shift in the answer distribution from the text-only baseline (Figure A.13) to the blind baseline, which demonstrates the impact caused by the captions alone. Moreover, we see the predicted answer distribution of MAPL zero-shot is closer to the ground truth answer distribution than that of the blind baseline (Figure A.14), which indicates that MAPL is leveraging the additional information from the visual input. For instance, we observe MAPL's predicted answer distribution for the "what color" question type (second column) looks more similar to the ground truth distribution compared to the text-only and blind baselines. Finally, when performing in-context learning from four shots (Figure A.16), we see the answer distribution gets even closer to the ground truth distribution. However, we do not observe much difference in answer distribution when increasing the number of shots from four to eight (Figure A.17).



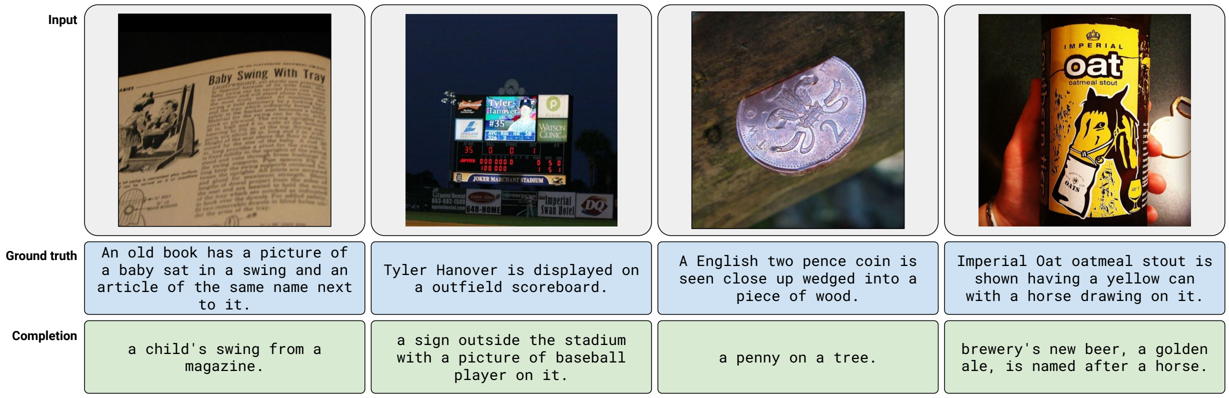
Ground truth	stadium coming alive on game day !.	and what 's the umbrella for ? a woman braves deep flood water.	my little brother will probably have this cake for his birthday.	the hand of person --.
Completion	sunset from the stadium.	a woman holds an umbrella as she walks through the flooded streets.	a cake for a video game themed birthday party.	a superman logo with a fist and lightning bolt.

Fig. A.1. MAPL's image captioning on Conceptual Captions.



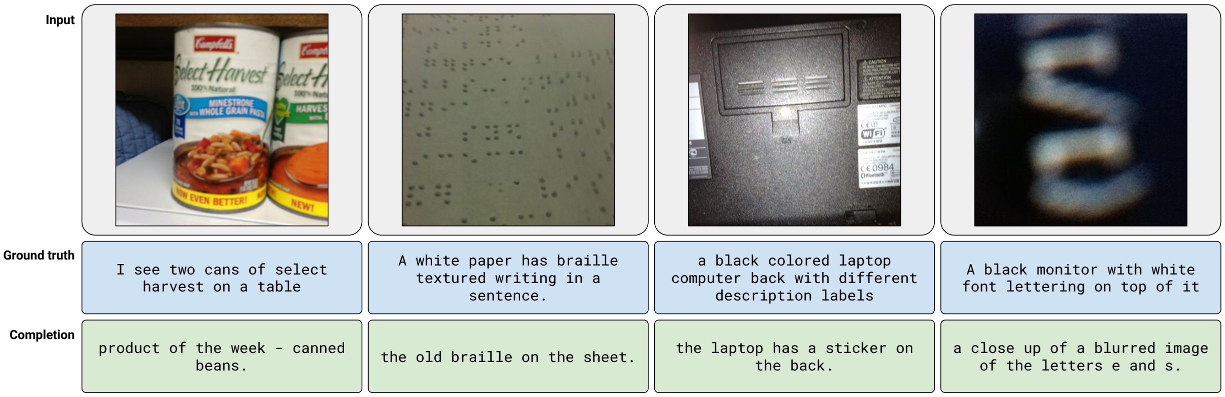
Ground truth	A couple of large long trains on a track.	A bird perched on top of a wooden power pole.	Parking meter that is expired, with damp street below	The meal consists of chopped chicken with cheese and brocolli.
Completion	a train passes a station with a woman and a child.	a crow perched on a power pole.	the vending machine is a relic of the past.	this is what i made for dinner tonight.

Fig. A.2. MAPL's image captioning on COCO Captions.



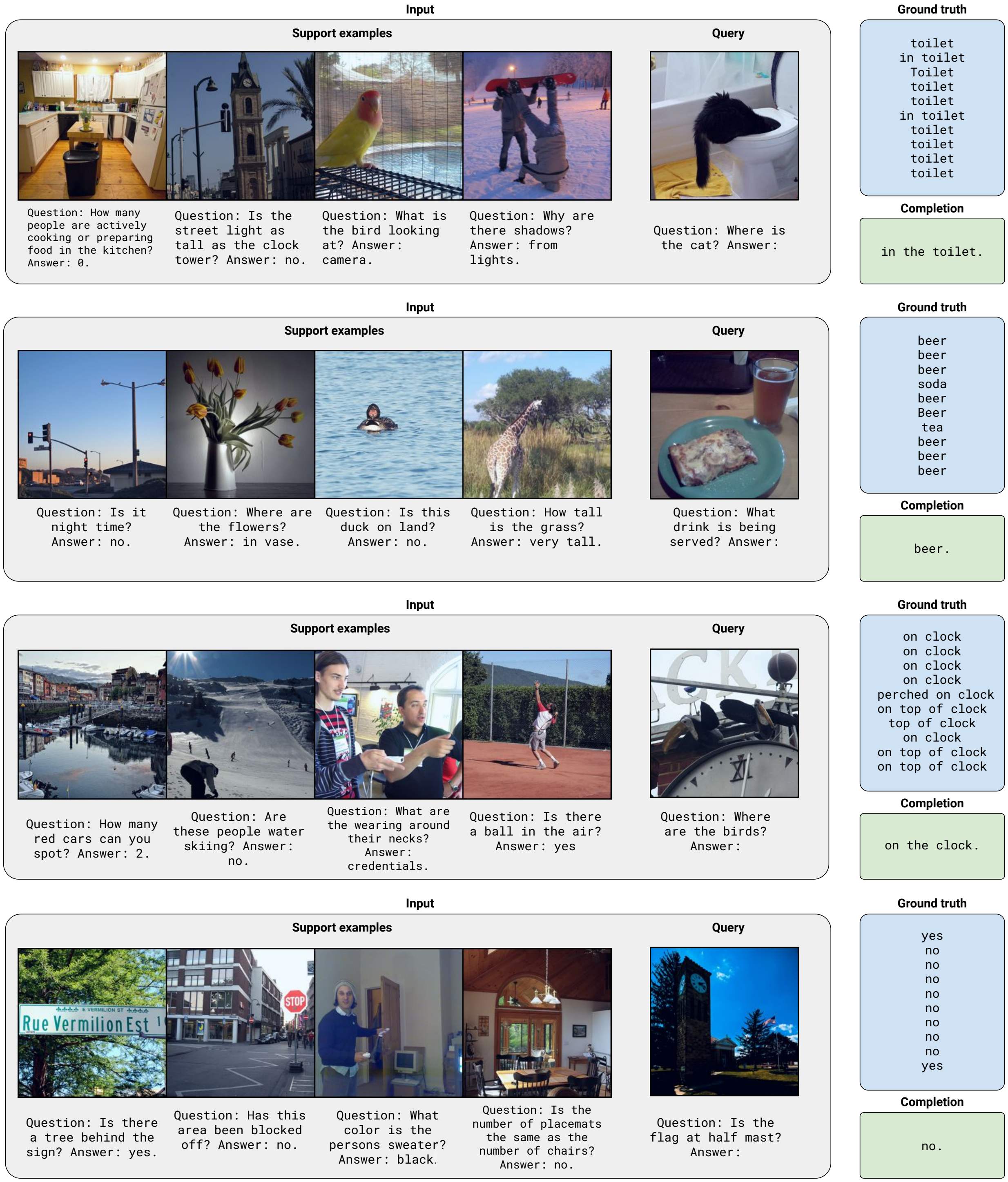
Ground truth	An old book has a picture of a baby sat in a swing and an article of the same name next to it.	Tyler Hanover is displayed on a outfield scoreboard.	A English two pence coin is seen close up wedged into a piece of wood.	Imperial Oat oatmeal stout is shown having a yellow can with a horse drawing on it.
Completion	a child's swing from a magazine.	a sign outside the stadium with a picture of baseball player on it.	a penny on a tree.	brewery's new beer, a golden ale, is named after a horse.

Fig. A.3. MAPL's image captioning on TextCaps.



Input				
Ground truth	I see two cans of select harvest on a table	A white paper has braille textured writing in a sentence.	a black colored laptop computer back with different description labels	A black monitor with white font lettering on top of it
Completion	product of the week - canned beans.	the old braille on the sheet.	the laptop has a sticker on the back.	a close up of a blurred image of the letters e and s.

Fig. A.4. MAPL's image captioning on VizWiz Captions.



Question: How many people are actively cooking or preparing food in the kitchen? Answer: 0.

Question: Is the street light as tall as the clock tower? Answer: no.

Question: What is the bird looking at? Answer: camera.

Question: Why are there shadows? Answer: from lights.

Question: Where is the cat? Answer:

toilet
in toilet
Toilet
toilet
toilet
in toilet
toilet
toilet
toilet
toilet

in the toilet.

Question: Is it night time? Answer: no.

Question: Where are the flowers? Answer: in vase.

Question: Is this duck on land? Answer: no.

Question: How tall is the grass? Answer: very tall.

Question: What drink is being served? Answer:

beer
beer
beer
soda
beer
Beer
tea
beer
beer
beer

beer.

Question: How many red cars can you spot? Answer: 2.

Question: Are these people water skiing? Answer: no.

Question: What are the wearing around their necks? Answer: credentials.

Question: Is there a ball in the air? Answer: yes

Question: Where are the birds? Answer:

on clock
on clock
on clock
on clock
perched on clock
on top of clock
top of clock
on clock
on top of clock
on top of clock

on the clock.

Question: Is there a tree behind the sign? Answer: yes.

Question: Has this area been blocked off? Answer: no.

Question: What color is the persons sweater? Answer: black.

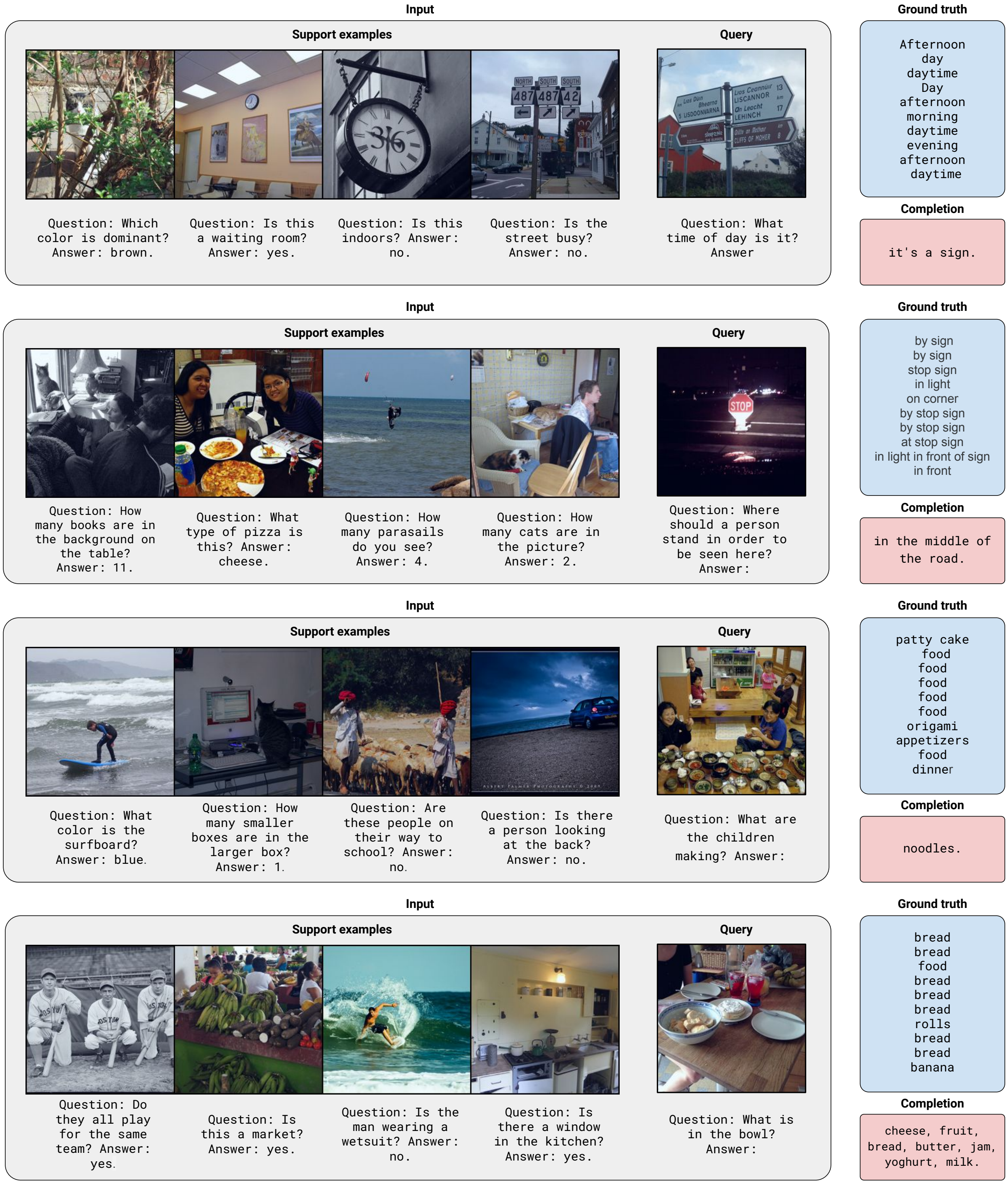
Question: Is the number of placemats the same as the number of chairs? Answer: no.

Question: Is the flag at half mast? Answer:

yes
no
no
no
no
no
no
no
no
yes

no.

Fig. A.5. MAPL's 4-shot VQA on VQAv2, success cases.



Input

Support examples

Question: Which color is dominant? Answer: brown.

Question: Is this a waiting room? Answer: yes.

Question: Is this indoors? Answer: no.

Question: Is the street busy? Answer: no.

Query

Question: What time of day is it? Answer

Ground truth

Afternoon
day
daytime
Day
afternoon
morning
daytime
evening
afternoon
daytime

Completion

it's a sign.

Input

Support examples

Question: How many books are in the background on the table? Answer: 11.

Question: What type of pizza is this? Answer: cheese.

Question: How many parasails do you see? Answer: 4.

Question: How many cats are in the picture? Answer: 2.

Query

Question: Where should a person stand in order to be seen here? Answer:

Ground truth

by sign
by sign
stop sign
in light
on corner
by stop sign
by stop sign
at stop sign
in light in front of sign
in front

Completion

in the middle of the road.

Input

Support examples

Question: What color is the surfboard? Answer: blue.

Question: How many smaller boxes are in the larger box? Answer: 1.

Question: Are these people on their way to school? Answer: no.

Question: Is there a person looking at the back? Answer: no.

Query

Question: What are the children making? Answer:

Ground truth

patty cake
food
food
food
food
food
origami
appetizers
food
dinner

Completion

noodles.

Input

Support examples

Question: Do they all play for the same team? Answer: yes.

Question: Is this a market? Answer: yes.

Question: Is the man wearing a wetsuit? Answer: no.

Question: Is there a window in the kitchen? Answer: yes.

Query

Question: What is in the bowl? Answer:

Ground truth

bread
bread
food
bread
bread
bread
rolls
bread
bread
banana

Completion

cheese, fruit, bread, butter, jam, yoghurt, milk.

Fig. A.6. MAPL's 4-shot VQA on VQAv2, failure cases.

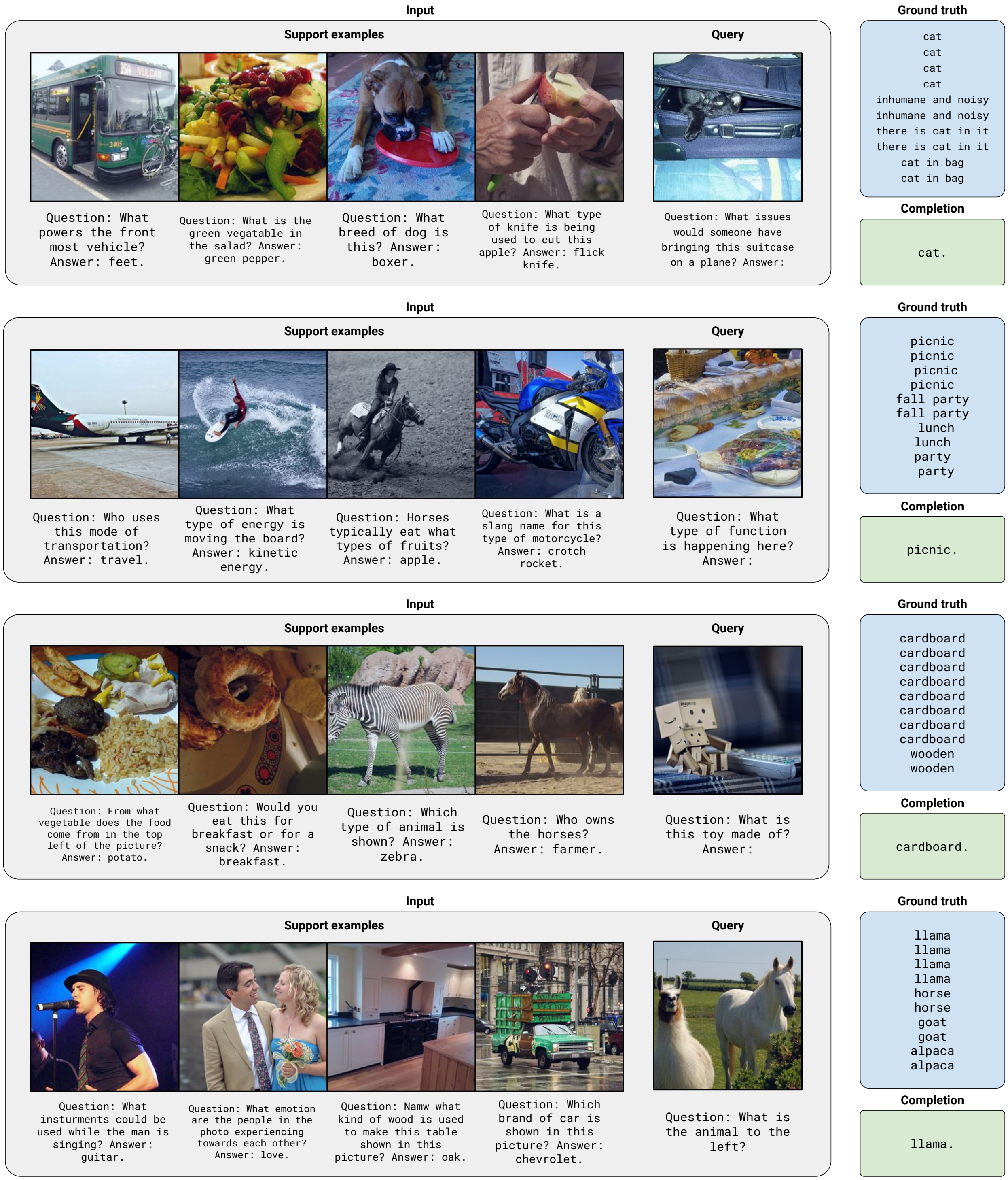
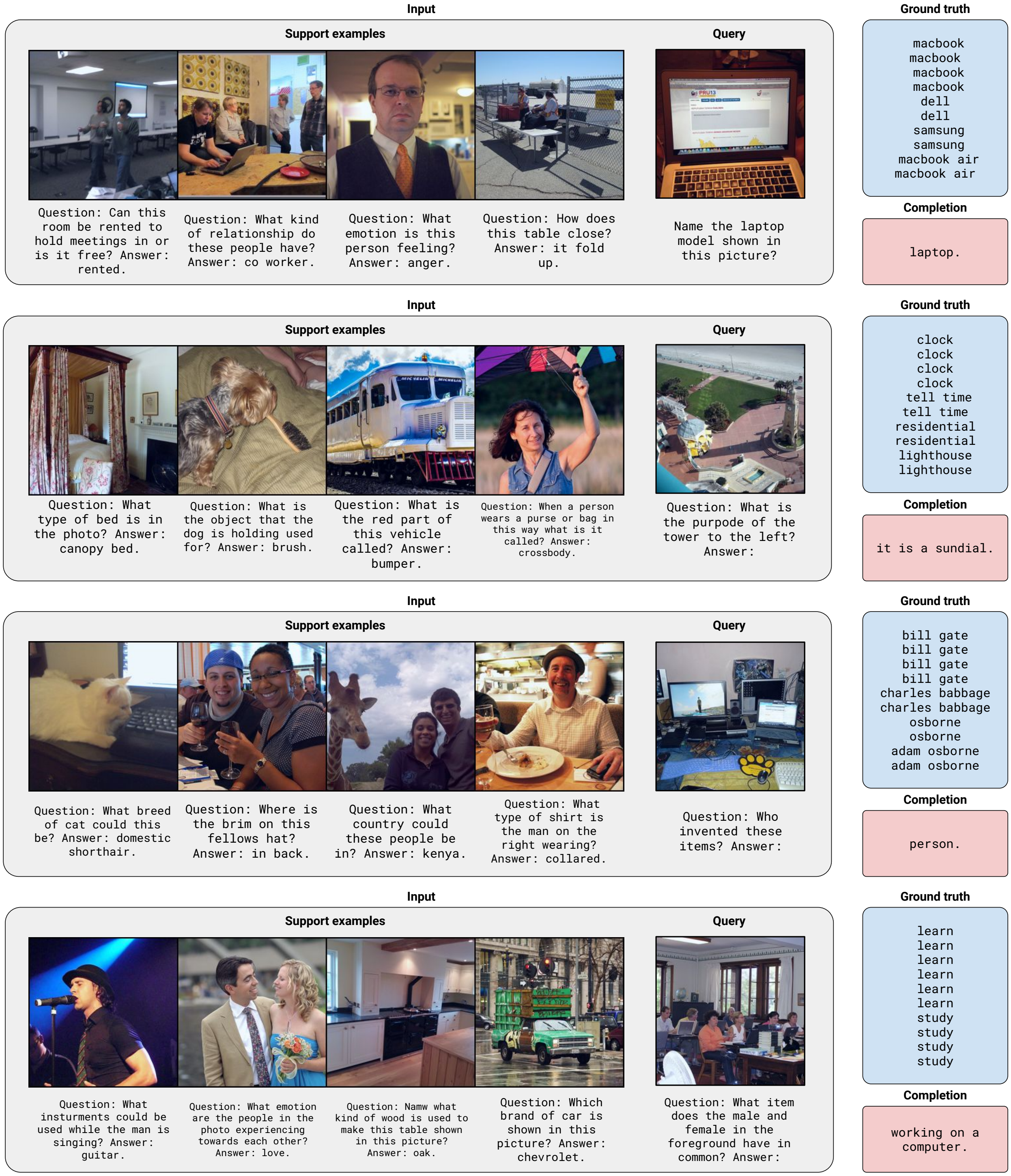

Input
Support examples
Query
Ground truth
cat
cat
cat
cat
inhumane and noisy
inhumane and noisy
there is cat in it
there is cat in it
cat in bag
cat in bag
Completion
cat.
Question: What powers the front most vehicle? Answer: feet.
Question: What is the green vegatable in the salad? Answer: green pepper.
Question: What breed of dog is this? Answer: boxer.
Question: What type of knife is being used to cut this apple? Answer: flick knife.
Question: What issues would someone have bringing this suitcase on a plane? Answer:
Input
Support examples
Query
Ground truth
picnic
picnic
picnic
picnic
fall party
fall party
lunch
lunch
party
party
Completion
picnic.
Question: Who uses this mode of transportation? Answer: travel.
Question: What type of energy is moving the board? Answer: kinetic energy.
Question: Horses typically eat what types of fruits? Answer: apple.
Question: What is a slang name for this type of motorcycle? Answer: crotch rocket.
Question: What type of function is happening here? Answer:
Input
Support examples
Query
Ground truth
cardboard
cardboard
cardboard
cardboard
cardboard
cardboard
cardboard
cardboard
wooden
wooden
Completion
cardboard.
Question: From what vegetable does the food come from in the top left of the picture? Answer: potato.
Question: Would you eat this for breakfast or for a snack? Answer: breakfast.
Question: Which type of animal is shown? Answer: zebra.
Question: Who owns the horses? Answer: farmer.
Question: What is this toy made of? Answer:
Input
Support examples
Query
Ground truth
llama
llama
llama
llama
horse
horse
goat
goat
alpaca
alpaca
Completion
llama.
Question: What insturments could be used while the man is singing? Answer: guitar.
Question: What emotion are the people in the photo experiencing towards each other? Answer: love.
Question: Namw what kind of wood is used to make this table shown in this picture? Answer: oak.
Question: Which brand of car is shown in this picture? Answer: chevrolet.
Question: What is the animal to the left?

Fig. A.7. MAPL's 4-shot VQA on OK-VQA, success cases.



Question: Can this room be rented to hold meetings in or is it free? Answer: rented.

Question: What kind of relationship do these people have? Answer: co worker.

Question: What emotion is this person feeling? Answer: anger.

Question: How does this table close? Answer: it fold up.

Name the laptop model shown in this picture?

macbook
macbook
macbook
macbook
dell
dell
samsung
samsung
macbook air
macbook air

laptop.

Question: What type of bed is in the photo? Answer: canopy bed.

Question: What is the object that the dog is holding used for? Answer: brush.

Question: What is the red part of this vehicle called? Answer: bumper.

Question: When a person wears a purse or bag in this way what is it called? Answer: crossbody.

Question: What is the purpode of the tower to the left? Answer:

clock
clock
clock
clock
tell time
tell time
residential
residential
lighthouse
lighthouse

it is a sundial.

Question: What breed of cat could this be? Answer: domestic shorthair.

Question: Where is the brim on this fellows hat? Answer: in back.

Question: What country could these people be in? Answer: kenya.

Question: What type of shirt is the man on the right wearing? Answer: collared.

Question: Who invented these items? Answer:

bill gate
bill gate
bill gate
bill gate
charles babbage
charles babbage
osborne
osborne
adam osborne
adam osborne

person.

Question: What insturments could be used while the man is singing? Answer: guitar.

Question: What emotion are the people in the photo experiencing towards each other? Answer: love.

Question: Namw what kind of wood is used to make this table shown in this picture? Answer: oak.

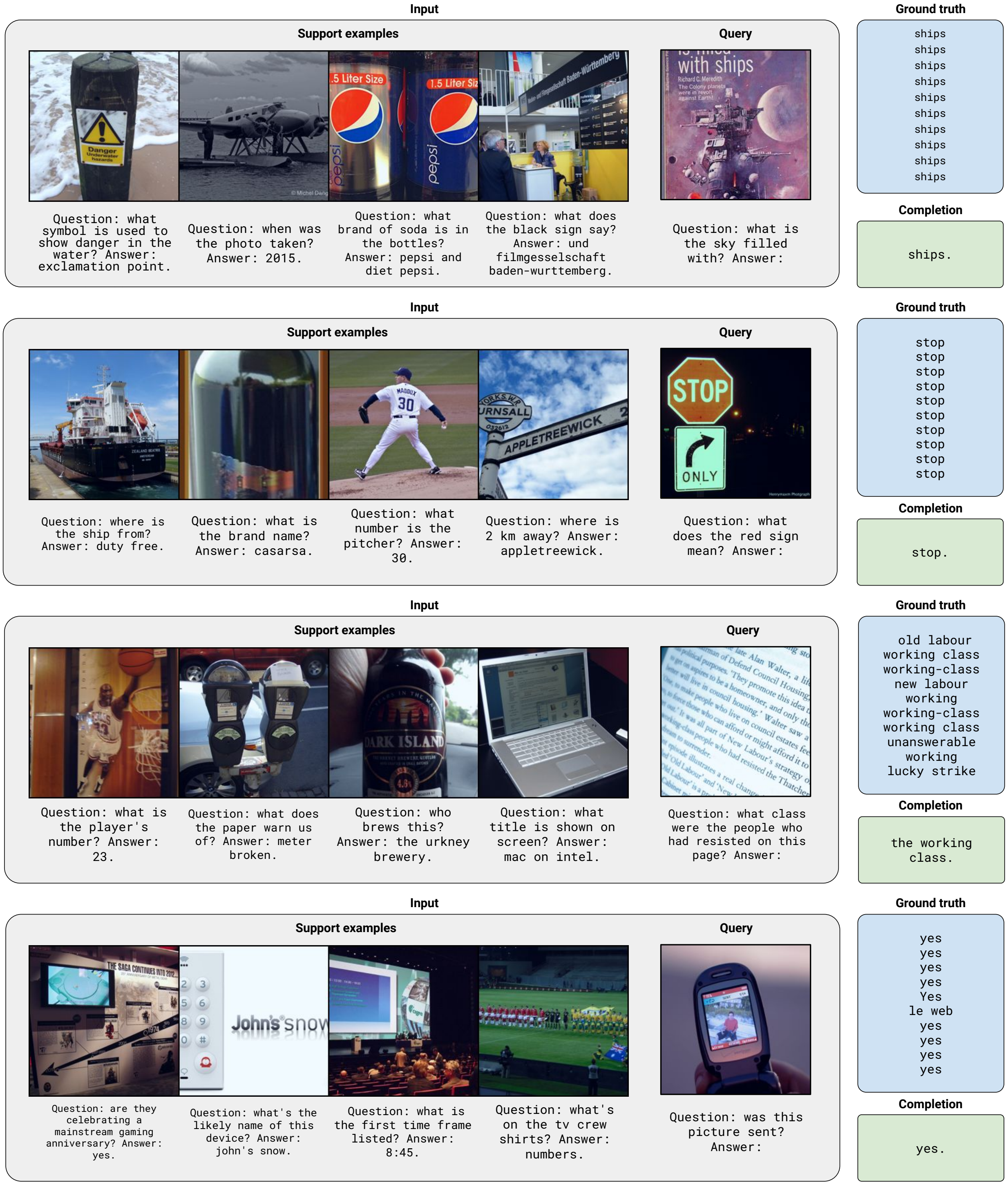
Question: Which brand of car is shown in this picture? Answer: chevrolet.

Question: What item does the male and female in the foreground have in common? Answer:

learn
learn
learn
learn
learn
learn
study
study
study
study

working on a computer.

Fig. A.8. MAPL's 4-shot VQA on OK-VQA, failure cases.



Question: what symbol is used to show danger in the water? Answer: exclamation point.

Question: when was the photo taken? Answer: 2015.

Question: what brand of soda is in the bottles? Answer: pepsi and diet pepsi.

Question: what does the black sign say? Answer: und filmgesselschaft baden-wurttemberg.

Question: what is the sky filled with? Answer:

ships
ships
ships
ships
ships
ships
ships
ships
ships
ships

ships.

Question: where is the ship from? Answer: duty free.

Question: what is the brand name? Answer: casarsa.

Question: what number is the pitcher? Answer: 30.

Question: where is 2 km away? Answer: appletreewick.

Question: what does the red sign mean? Answer:

stop
stop
stop
stop
stop
stop
stop
stop
stop
stop

stop.

Question: what is the player's number? Answer: 23.

Question: what does the paper warn us of? Answer: meter broken.

Question: who brews this? Answer: the urkney brewery.

Question: what title is shown on screen? Answer: mac on intel.

Question: what class were the people who had resisted on this page? Answer:

old labour
working class
working-class
new labour
working
working-class
working class
unanswerable
working
lucky strike

the working class.

Question: are they celebrating a mainstream gaming anniversary? Answer: yes.

Question: what's the likely name of this device? Answer: john's snow.

Question: what is the first time frame listed? Answer: 8:45.

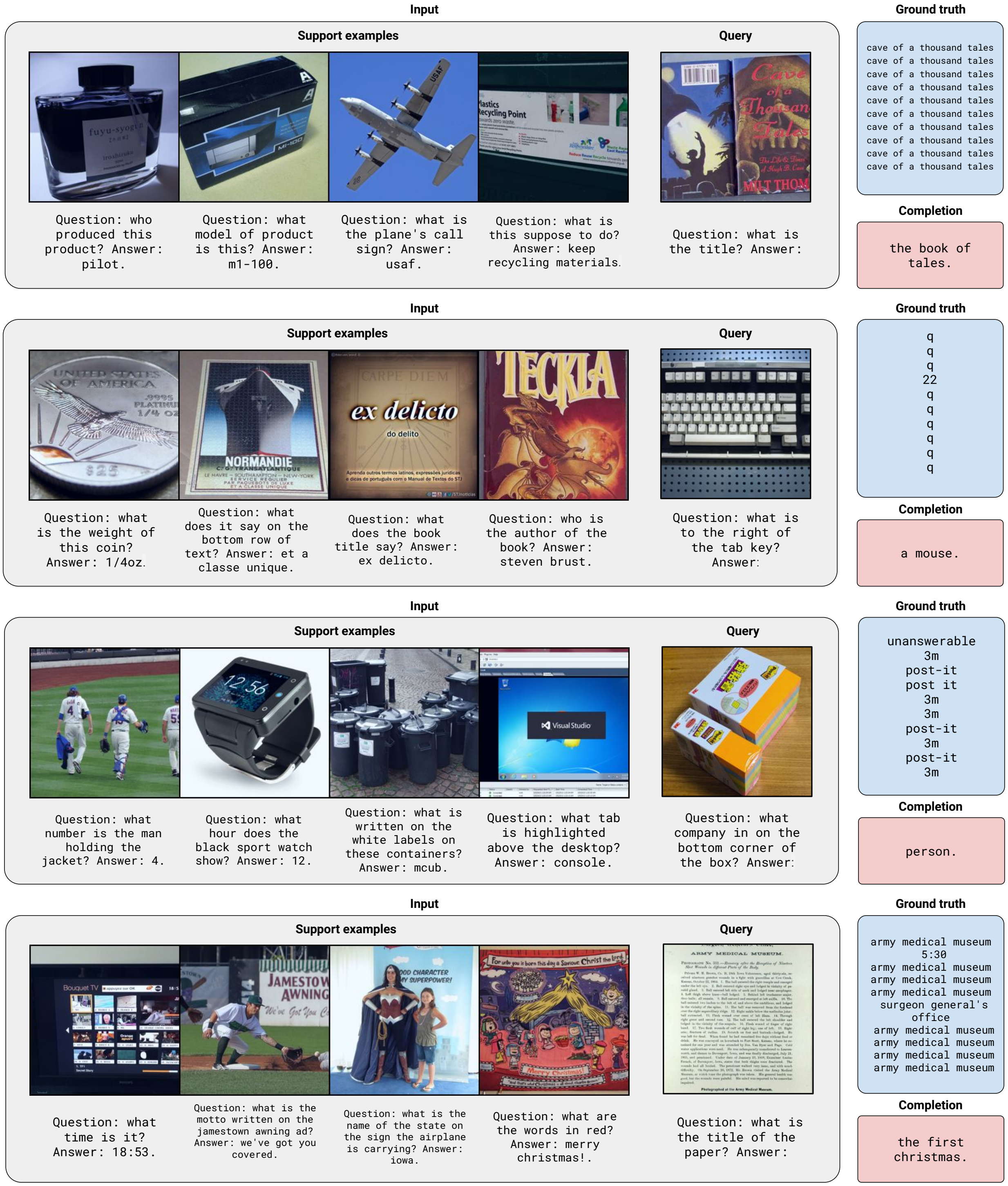
Question: what's on the tv crew shirts? Answer: numbers.

Question: was this picture sent? Answer:

yes
yes
yes
yes
Yes
le web
yes
yes
yes
yes

yes.

Fig. A.9. MAPL's 4-shot VQA on TextVQA, success cases.



Question: who produced this product? Answer: pilot.

Question: what model of product is this? Answer: m1-100.

Question: what is the plane's call sign? Answer: usaf.

Question: what is this suppose to do? Answer: keep recycling materials.

Question: what is the title? Answer:

cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales
cave of a thousand tales

the book of tales.

Question: what is the weight of this coin? Answer: 1/4oz.

Question: what does it say on the bottom row of text? Answer: et a classe unique.

Question: what does the book title say? Answer: ex delicto.

Question: who is the author of the book? Answer: steven brust.

Question: what is to the right of the tab key? Answer:

q
q
q
22
q
q
q
q
q
q

a mouse.

Question: what number is the man holding the jacket? Answer: 4.

Question: what hour does the black sport watch show? Answer: 12.

Question: what is written on the white labels on these containers? Answer: mcub.

Question: what tab is highlighted above the desktop? Answer: console.

Question: what company in on the bottom corner of the box? Answer:

unanswerable
3m
post-it
post it
3m
3m
post-it
3m
post-it
3m

person.

Question: what time is it? Answer: 18:53.

Question: what is the motto written on the jamestown awning ad? Answer: we've got you covered.

Question: what is the name of the state on the sign the airplane is carrying? Answer: iowa.

Question: what are the words in red? Answer: merry christmas!.

Question: what is the title of the paper? Answer:

army medical museum
5:30
army medical museum
army medical museum
army medical museum
surgeon general's office
army medical museum
army medical museum
army medical museum
army medical museum

the first christmas.

Fig. A.10. MAPL's 4-shot VQA on TextVQA, failure cases.

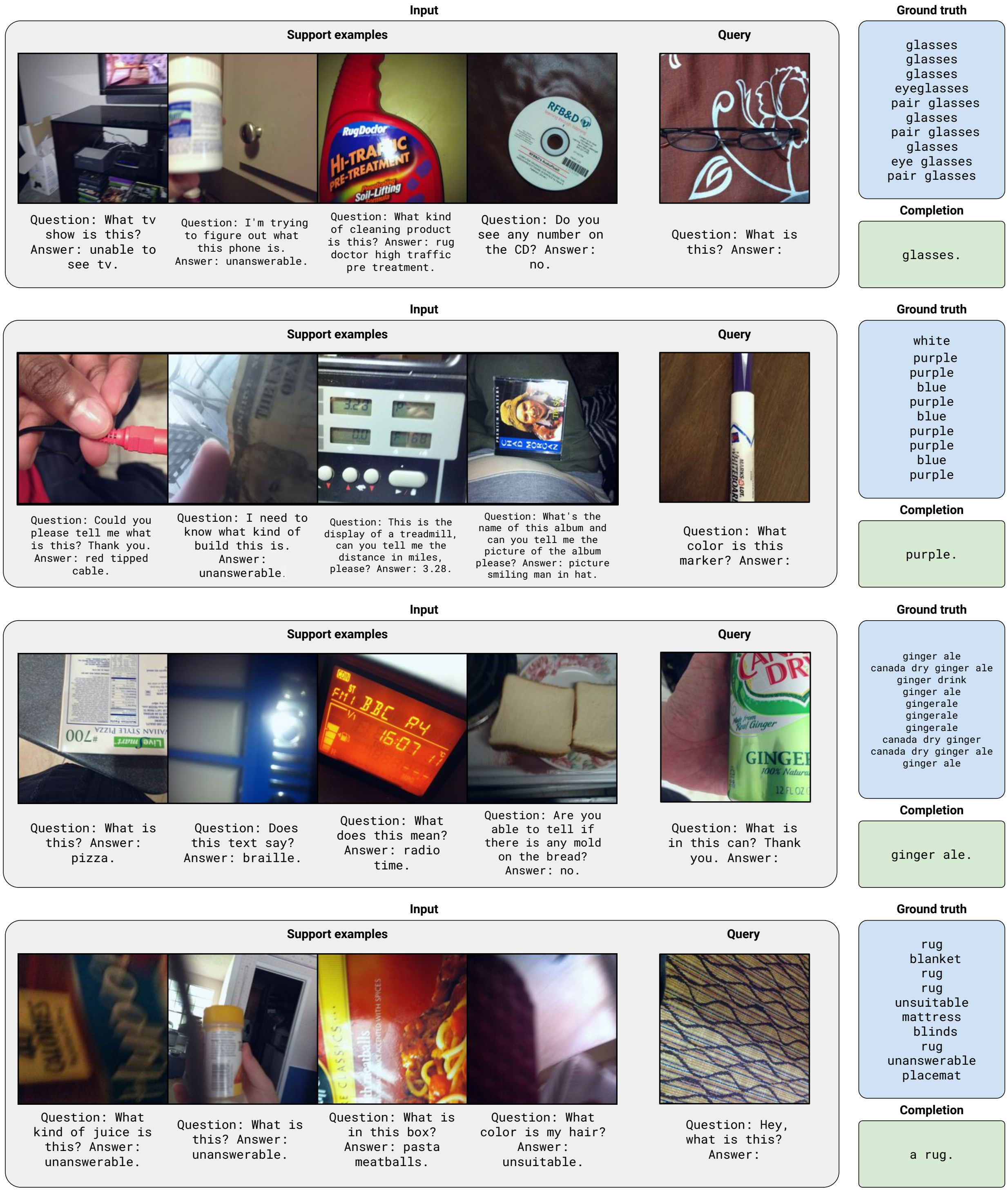
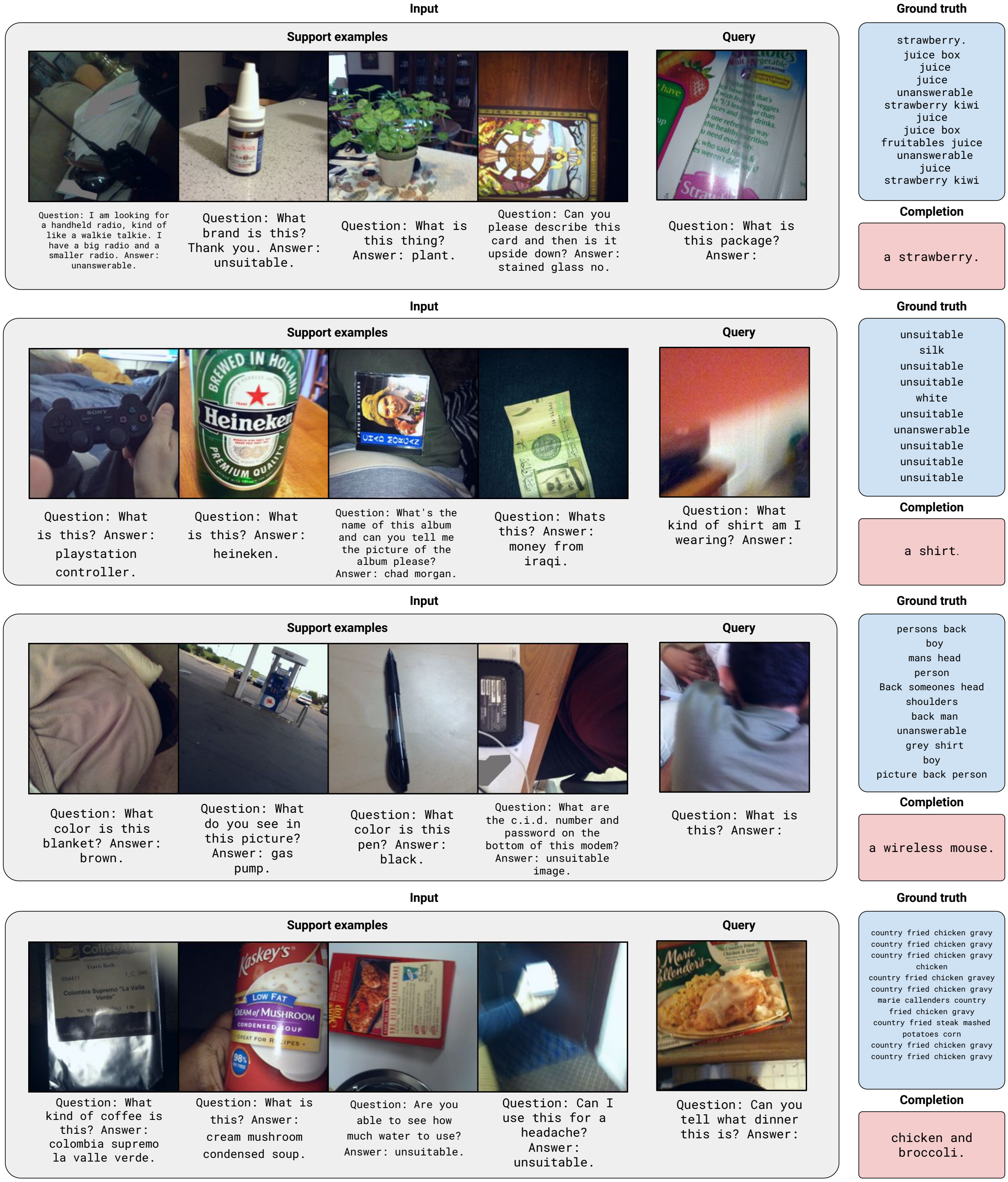

Input
Support examples
Query
Ground truth
Question: What tv show is this? Answer: unable to see tv.
Question: I'm trying to figure out what this phone is. Answer: unanswerable.
Question: What kind of cleaning product is this? Answer: rug doctor high traffic pre treatment.
Question: Do you see any number on the CD? Answer: no.
Question: What is this? Answer:
glasses
glasses
glasses
eyeglasses
pair glasses
glasses
pair glasses
glasses
eye glasses
pair glasses
Completion
glasses.
Input
Support examples
Query
Ground truth
Question: Could you please tell me what is this? Thank you. Answer: red tipped cable.
Question: I need to know what kind of build this is. Answer: unanswerable.
Question: This is the display of a treadmill, can you tell me the distance in miles, please? Answer: 3.28.
Question: What's the name of this album and can you tell me the picture of the album please? Answer: picture smiling man in hat.
Question: What color is this marker? Answer:
white
purple
purple
blue
purple
blue
purple
purple
blue
purple
Completion
purple.
Input
Support examples
Query
Ground truth
Question: What is this? Answer: pizza.
Question: Does this text say? Answer: braille.
Question: What does this mean? Answer: radio time.
Question: Are you able to tell if there is any mold on the bread? Answer: no.
Question: What is in this can? Thank you. Answer:
ginger ale
canada dry ginger ale
ginger drink
ginger ale
gingerale
gingerale
gingerale
canada dry ginger
canada dry ginger ale
ginger ale
Completion
ginger ale.
Input
Support examples
Query
Ground truth
Question: What kind of juice is this? Answer: unanswerable.
Question: What is this? Answer: unanswerable.
Question: What is in this box? Answer: pasta meatballs.
Question: What color is my hair? Answer: unsuitable.
Question: Hey, what is this? Answer:
rug
blanket
rug
rug
unsuitable
mattress
blinds
rug
unanswerable
placemat
Completion
a rug.

Fig. A.11. MAPL's 4-shot VQA on VizWiz-VQA, success cases.



Question: I am looking for a handheld radio, kind of like a walkie talkie. I have a big radio and a smaller radio. Answer: unanswerable.

Question: What brand is this? Thank you. Answer: unsuitable.

Question: What is this thing? Answer: plant.

Question: Can you please describe this card and then is it upside down? Answer: stained glass no.

Question: What is this package? Answer:

strawberry.
juice box
juice
juice
unanswerable
strawberry kiwi
juice
juice box
fruitables juice
unanswerable
juice
strawberry kiwi

a strawberry.

Question: What is this? Answer: playstation controller.

Question: What is this? Answer: heineken.

Question: What's the name of this album and can you tell me the picture of the album please? Answer: chad morgan.

Question: Whats this? Answer: money from iraqi.

Question: What kind of shirt am I wearing? Answer:

unsuitable
silk
unsuitable
unsuitable
white
unsuitable
unanswerable
unsuitable
unsuitable
unsuitable

a shirt.

Question: What color is this blanket? Answer: brown.

Question: What do you see in this picture? Answer: gas pump.

Question: What color is this pen? Answer: black.

Question: What are the c.i.d. number and password on the bottom of this modem? Answer: unsuitable image.

Question: What is this? Answer:

persons back
boy
mans head
person
Back someones head
shoulders
back man
unanswerable
grey shirt
boy
picture back person

a wireless mouse.

Question: What kind of coffee is this? Answer: colombia supremo la valle verde.

Question: What is this? Answer: cream mushroom condensed soup.

Question: Are you able to see how much water to use? Answer: unsuitable.

Question: Can I use this for a headache? Answer: unsuitable.

Question: Can you tell what dinner this is? Answer:

country fried chicken gravy
country fried chicken gravy
country fried chicken gravy
chicken
country fried chicken gravey
country fried chicken gravy
marie callenders country fried chicken gravy
country fried steak mashed potatoes corn
country fried chicken gravy
country fried chicken gravy

chicken and broccoli.

Fig. A.12. MAPL's 4-shot VQA on VizWiz-VQA, failure cases.

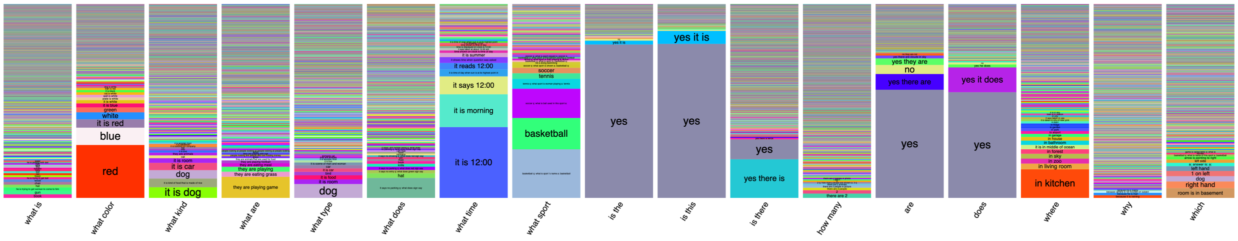


Fig. A.13. Predicted answer distributions for selected VQAv2 question types with the text-only baseline.

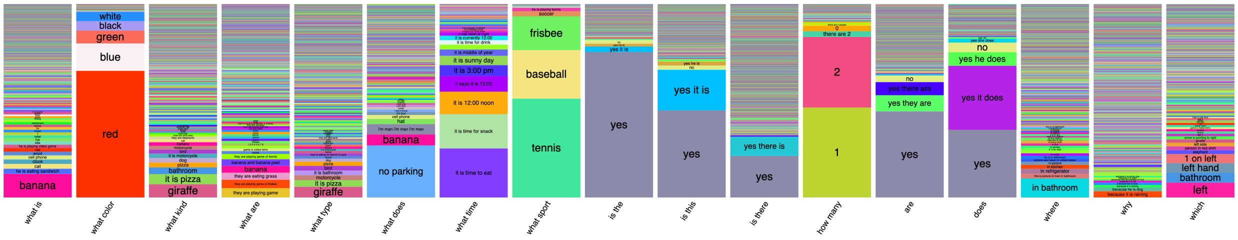


Fig. A.14. Predicted answer distributions for selected VQAv2 question types with the blind baseline.

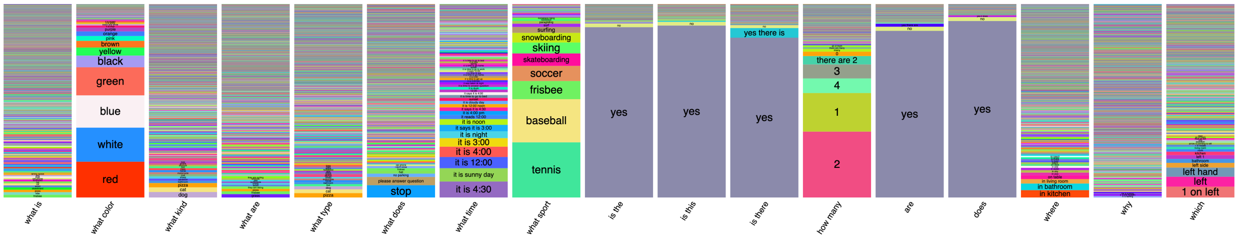


Fig. A.15. Predicted answer distributions for selected VQAv2 question types with MAPL 0-shot.

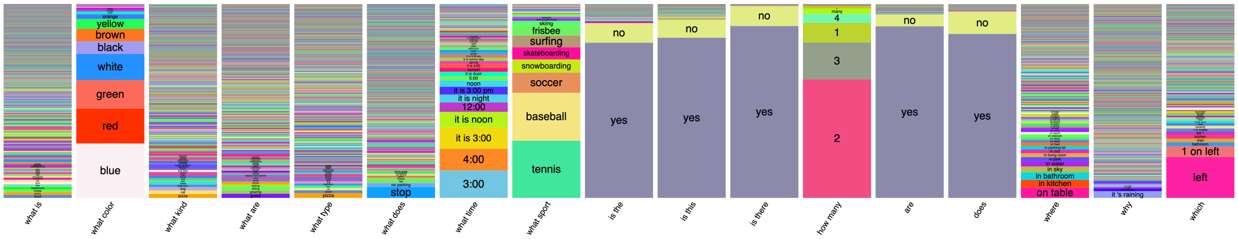


Fig. A.16. Predicted answer distributions for selected VQAv2 question types with MAPL 4-shot.

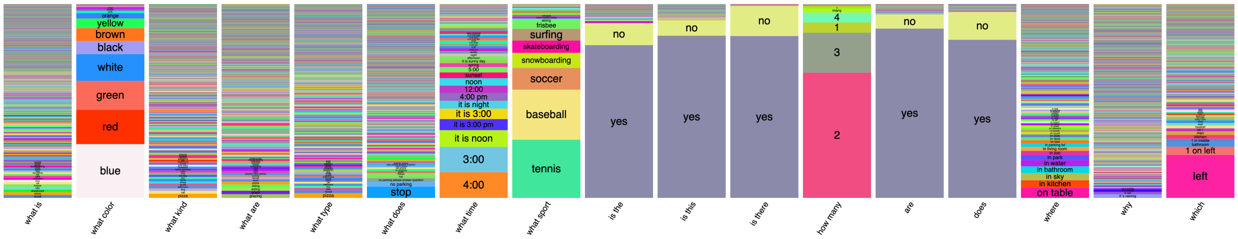


Fig. A.17. Predicted answer distributions for selected VQAv2 question types with MAPL 8-shot.

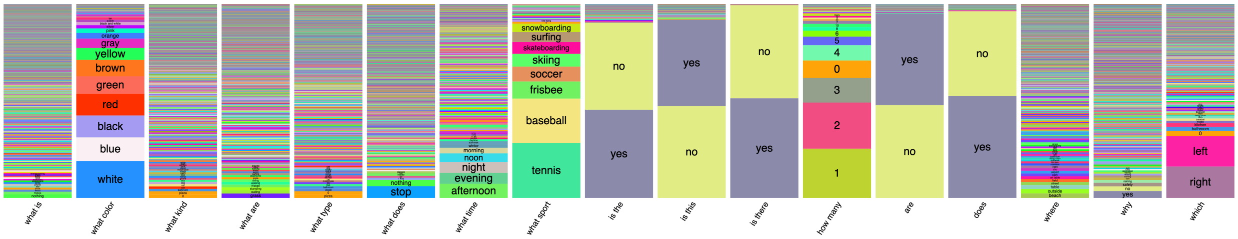


Fig. A.18. Ground truth answer distributions for selected VQAv2 question types.

Appendix B

Appendix for An Examination of the Robustness of Reference-Free Image Captioning Evaluation Metrics

B.0.1. Generating Caption-like Sentences

To generate caption-like sentences from each question and answer pair of VQA datasets, we utilize pre-trained GPT-J [**100**] in a few-shot manner. To accomplish this, we first constructed a support examples dataset using the VQAv2 [**32**] training split. For each of the sixty-four predefined question types in the VQAv2 dataset, we randomly selected four examples from the VQAv2 training split. Then, we transformed both the questions and answers into single sentences, which we wrote ourselves. When generating captions for VQAv2 validation split, we first match the question type to one of the predefined sixty-four question types. Then, we select four support examples associated with that question type and prompt GPT-J to generate a transformed sentence. If the question type does not match any of our predefined question types, we randomly select eight support examples from the entire pool of support examples. Please see Figure 3.2 and note that we visualized a 2-shot prompt for simplification.

B.0.2. F1 score computation for the Composite Dataset

We calculated the F1 score between the human-written correct captions and model generated incorrect captions in the Composite dataset [**1**]. We used the captions generated by the Karparthy model [**45**] as they were better in quality. In the Composite dataset, each model generated caption has an associated correctness score (provided by humans) ranging from 1 ('The description has no relevance to the image') to 5 ('The description relates perfectly to the image'). For our F1 score computation, we considered all captions with score less than or equal to 4 as incorrect captions.

B.0.3. Plausible Answers

To generate plausible captions for each question type, we first compiled a list of plausible answers derived from the ground truth multiple-choice answer of the same question type in the validation split of VQAv2. Subsequently, an answer was randomly selected from this list of plausible answers. This chosen answer was used to replace the ground truth answer in the original caption, thus generating a plausible alternate caption.

B.0.4. Picking a large and small object from the image

In this experiment, our primary objective is to investigate how the object size mentioned in captions affects the scores assigned by CLIPScore and UMIC. To select small and large objects that are distinctly different in size, we could sort the objects by their associated area in the COCO Detection dataset. However, this approach may not always yield accurate results because multiple objects with the same name may appear in an image. For instance, if there are two cars in an image, one smaller but further away and the other larger but closer, sorting by area would lead to incorrect identification of the smallest and largest objects. This would result in identical captions for both objects, such as “There is a car." which is not ideal for comparison.

To overcome this issue, we added up the area of all object categories with the same name and sorted the total areas of each object category in the image. We then calculated the difference between the areas associated with the largest and smallest categories. If the difference exceeded our threshold, we selected those objects for analysis. As a result, we selected 24610 images for further analysis (See Figure 3.6).

B.0.5. Computational Resources

In all experiments detailed in this paper, we employed a single NVIDIA Quadro RTX 8000 with 48 GB GDDR6 GPU Memory. Specifically, for the primary task of generating caption-like sentences from the VQAv2 dataset, we performed inference using the GPT-J model with 6 billion parameters, executing the process over a duration of 24 hours.

B.0.6. Dataset Terms of Use

We will distribute our datasets (both generated with caption template and QA to caption conversion method) under the Creative Commons Attribution 4.0 License. It is noteworthy to mention that this licensing choice aligns with the terms of use governing both the COCO and VQAv2 datasets, foundational to the creation of our datasets.

B.0.7. Editorial Assistance

We would like to disclose that ChatGPT was utilized for refining the language and structure of this academic paper. While the primary content and research remain the work of the authors, the assistance provided by ChatGPT was limited to the improvement of writing quality.