Enhancing Compositional Reasoning of Vision-Language Models

Geerisha Jain* **Madhura Deshpande*** **Shaurya Singh*** **Shrey Madeyanda***
{geerishj, mvdeshpa, shauryas, smadeyan}@andrew.cmu.edu

Abstract

Vision-Language Models (VLMs) are used for a wide variety of tasks that benefit from compositional reasoning. However, in the image-to-text retrieval context, current state-of-the-art models often have trouble binding attributes in the text to the corresponding objects in a manner that correctly reflects the image. We propose to improve the compositional reasoning of VLMs by focusing on attribute binding within the text modality, to emphasize the object-attribute relationships in the text, followed by aligning the resulting text representation with the image representation using cross-modal attention mechanism. Finally, we aim to learn a shared representation for the two modalities using representation fusion, which will be more useful than individual representations of the two modalities.

1 Introduction and Problem Definition

Compositionality, the understanding that "the meaning of the whole is a function of its parts" (Chen et al., 2020), is an important aspect of intelligence. In natural language, a sentence is made up of its words. For vision, we can consider an image, which is made up of parts like objects, their attributes, and their relationships (Hudson and Manning, 2019) (**?**). For instance, compositionality allows people to differentiate between a photo of "a man in a yellow shirt facing a wall painted white" and "a man in a white shirt facing a wall painted yellow".

Today's vision-language models, pretrained on large-scale image-caption datasets, are being widely applied for tasks that benefit from compositional reasoning, including retrieval, text-to-image generation, and open-vocabulary classification (Ma et al., 2023). Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) struggle with binding correct attributes to the correct objects, understanding relations between objects, generalizing systematically to unseen combinations of concepts and to larger and more complex sentences (Singh et al., 2023). We aim to implement an approach that improves the compositional reasoning of vision language models through a focus on attribute binding.

In the context of vision-language models and compositionality, attribute binding refers to the process of correctly associating descriptive attributes (such as colors, shapes, sizes, etc.) with their corresponding objects or entities depicted in an image. This task involves understanding the spatial relationships between objects and their attributes as described in textual descriptions. For example, in the sentence "a red apple on a table," attribute binding involves correctly associating the attribute "red" with the object "apple" and understanding the spatial relationship between the apple and the table.

Attribute binding is crucial for generating accurate and meaningful descriptions of visual scenes in natural language. It requires the model to not only recognize objects and their attributes within an image but also to understand how these attributes are related to the objects and their spatial configurations. Achieving effective attribute binding contributes to the overall compositional capability of vision-language models, allowing them to generate coherent and contextually relevant descriptions of complex visual scenes.

Current state-of-the-art models like CLIP, which utilize a contrastive learning approach, lack the fine-grained multimodal alignment necessary for complex compositional reasoning tasks. This limitation hinders the models' ability to capture the

*Everyone Contributed Equally – Alphabetical order

intricate object-attribute relationships that are crucial for precise image understanding and retrieval.

To address this challenge, we propose an approach that enhances the compositional reasoning capabilities of VLMs by explicitly modeling attribute binding within the text modality. Our method emphasizes the object-attribute relationships in the text representation, ensuring that the attributes are tightly coupled with their respective objects, using language-driven cross attention losses proposed in (Rassin et al., 2024). Subsequently, we employ a cross-modal attention mechanism to align the resulting text representation with the image representation, facilitating a more accurate and coherent understanding of the visual and textual information.

Through our approach, we aim to improve the attribute binding performance of VLMs, enabling them to more effectively capture the nuanced relationships between objects and their associated attributes in both text and image domains. This enhanced compositional reasoning capability is expected to boost the performance of VLMs in image-to-text retrieval tasks, as well as other applications that require precise understanding of object-attribute relationships.

2 Related Work and Background

2.1 Related Datasets

2.1.1 SUGARCREPE

Recent benchmarks such as Winoground, VL-CheckList, ARO, CREPE, and Cola have emerged to evaluate the compositional reasoning capabilities of vision-language models through image-to-text retrieval tasks. However, a critical vulnerability has been uncovered across these benchmarks: significant biases render them hackable, with blind models outperforming state-of-the-art vision-language models. To address this issue, the paper (Hsieh et al., 2023) introduces SUGARCREPE, a novel benchmark designed to evaluate compositionality more faithfully.

Leveraging large language models, SUGARCREPE generates natural and plausible hard negative examples, overcoming the limitations of previous benchmarks relying on rule-based templates. Additionally, an adversarial refinement process is employed to mitigate biases, resulting in a more balanced dataset. Through both qualitative and quantitative evaluations, the effectiveness of SUGARCREPE in fixing biases is demonstrated.

The paper also re-evaluates recent methods aimed at improving compositionality, highlighting the overestimation of certain approaches when evaluated on existing benchmarks and suggesting the need for more innovative techniques to enhance compositional understanding in vision-language models.

2.1.2 MS COCO

The Microsoft Common Objects in Context (MS-COCO) dataset proposed in (Lin et al., 2015) is a large-scale, high-quality dataset for object detection, segmentation, and captioning tasks in computer vision. It is a widely adopted benchmark for evaluating and comparing the performance of various deep learning models in these tasks. The dataset contains over 300,000 images, with each image annotated with object bounding boxes, segmentation masks, and captions describing the scene.

2.1.3 ARO benchmark

(Yuksekgonul et al., 2023) identify significant gaps in vision-language models' abilities to understand relationships, attribution, and order within data, despite these models performing adequately on cross-modal retrieval tasks. They demonstrate that models can achieve a high performance on retrieval even when the order and composition cues are removed from captions or images, which leads to models with compositional deficiencies still performing well on the standard evaluations.

To address this concern, the authors introduced the Attribution, Relation, and Order benchmark (ARO) for fine-grained evaluation of VLMs' relation, attribution, and order understanding. They also propose introducing hard negatives consisting of the nearest neighboring images into each batch during training these models, to force models to represent fine-grained differences between very similar scenes.

2.1.4 VALSE benchmark

(Parcalabescu et al., 2022) introduce Vision And Language Structured Evaluation (VALSE), a novel benchmark designed to evaluate the visio-linguistic grounding capabilities of pretrained VLMs. Unlike traditional task-specific evaluations, VALSE focuses on specific linguistic phenomena. The authors emphasize that existing models struggle with

handling various linguistic constructs, and there is a need for a more nuanced assessment. VALSE comprises of six distinct tests, each targeting different aspects of linguistic understanding. These tests challenge models to ground linguistic expressions in visual context, covering areas such as object naming, spatial relations, and quantification.

By providing fine-grained evaluations, VALSE aims to complement existing benchmarks and drive progress in pretrained VLMs from a linguistic perspective. Moreover, VALSE's focus on grounding linguistic expressions in visual scenes highlights the need for models to handle contextual nuances effectively.

2.2 Unimodal Baselines

2.2.1 Vera

In this paper, (Liu et al., 2023) the authors introduce VERA, a general-purpose model designed for the verification of commonsense plausibility in declarative statements. The development of VERA responds to the problem of commonsense errors in the outputs of advanced language models. It integrates a novel training regimen that emphasizes the distinction between correct and incorrect assertions using a combination of binary classification, multi-class classification, and supervised contrastive training objectives.

VERA demonstrates substantial improvements over existing models, including the latest iterations of GPT models, by accurately verifying the plausibility of commonsense assertions. It significantly enhances the reliability of model-generated text, offering a robust method for the retrospective analysis of commonsense validity. The model's generalization abilities are highlighted through its performance on unseen data and its calibrated confidence levels, ensuring its applicability in real-world settings.

This framework is particularly relevant in scenarios where commonsense reasoning is crucial, providing a critical layer of validation to the outputs of generative language models.

2.2.2 An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

In the paper (Dosovitskiy et al., 2021), the authors propose a model architecture that adapts the transformer architecture, which had achieved remarkable success in natural language processing, to the domain of computer vision for image recognition tasks.

The authors proposed a novel method of treating an image as a sequence of flattened image patches, or "visual tokens." These visual tokens were then linearly projected into lower-dimensional representations and processed by a standard transformer encoder, enabling the model to capture long-range dependencies and global context across the entire image, a capability that traditional CNNs with local receptive fields struggle to achieve.

The authors evaluated their Vision Transformer (ViT) model on several large-scale image recognition benchmarks, such as ImageNet, and demonstrated competitive performance compared to state-of-the-art CNN-based models, despite being trained on significantly less data. This highlights the potential of transformers to be highly data-efficient and generalize well across various computer vision tasks. The ViT's ability to model global relationships between image patches proved to be a key advantage over traditional CNNs.

2.3 Prior Work

2.3.1 Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation

The Reinforced Cross-Modal (RCM) framework, proposed by (Wang et al., 2019) enhances cross-modal grounding through a dual focus on local visual scenes and global instruction-trajectory alignment, using reinforcement learning for dynamic interaction and a matching critic for intrinsic rewards based on instruction adherence. Self-Supervised Imitation Learning (SIL) is introduced to improve navigation in unseen environments by allowing an agent to imitate its successful past actions, thus refining its decision-making process over time. The experimental results indicate that the RCM approach outperforms existing methods, achieving new state-of-the-art results on the Room-to-Room dataset. The integration of SIL further enhances performance by enabling the model to generalize better to new environments, (Wang et al., 2019) Shrey

2.3.2 Linearly Mapping from Image to Text Space

In this paper, (Merullo et al., 2023) explores the transfer of visual concepts from images to language models using different image encoders with varying levels of linguistic supervision. They introduce a method called LiMBeR (Linearly Mapping Between Representation spaces), which linearly projects image representations into the input space of language models without modifying other model parameters. This approach is used to evaluate how pretrained language models can generate text-based descriptions from visual inputs, emphasizing the structural similarities between language and vision representation spaces.

Their evaluation show that Encoders with linguistic supervision (CLIP, NFRN50) are better at translating complex visual scenes into accurate textual descriptions compared to BEIT, which only manages coarse-grained visual properties. This indicates that linguistically informed encoders are more adept at compositional reasoning, aligning visual properties with the correct lexical categories .

2.3.3 Iterated Learning Improves Compositionality in Large Vision-Language Models

In this paper, (Zheng et al., 2024) tackles a fundamental challenge shared by both human vision and natural language: compositionality. Despite the performance gains achieved by large vision and language pretraining, recent investigations reveal that state-of-the-art VLMs struggle with compositionality. Further, they find that scaling up models or increasing training data doesn't guarantee improved compositional reasoning, challenging the underlying assumption that larger models inherently lead to better understanding of complex relationships in vision and language.

To address these challenges, the authors introduce an iterated training algorithm inspired by cognitive science research. This paradigm mimics the process of cultural transmission, where knowledge is passed down through generations. By iteratively resetting one of the agent's weights during training, the model induces representations that become "easier to learn." This approach encourages the emergence of compositional features, bridging the gap between human-like understanding and machine learning representations. They further propose a novel approach that combines graph decomposition and augmentation with negative mining techniques in the scene graph space. By leveraging scene graphs as a proxy for understanding image compositionality, their structured framework encourages the model to capture relationships between objects, attributes, and linguistic constructs. Additionally, carefully selected negative examples during training enhance attribute binding and relation understanding, ultimately improving downstream task performance. In summary, this work not only identifies the challenges related to compositionality but also proposes innovative techniques, such as iterated learning and scene graph-based representations, to address these limitations in large VLMs

2.3.4 Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts

In this paper, (Mustafa et al., 2022) introduces LIMoE, a multimodal sparse mixture of experts model capable of processing both images and text through a shared architecture. Utilizing a mixture of experts, LIMoE leverages sparsely-activated models to handle different data types concurrently, enhancing training efficiency and computational cost-effectiveness. The model is trained using a contrastive learning framework, which aligns the representations of images and text, thereby overcoming challenges like training stability and expert utilization imbalance.

Significant contributions include the introduction of an entropy-based regularization scheme that ensures balanced expert utilization and stabilizes training across varying scales of model architectures. LIMoE significantly outperforms dense models of equivalent computational cost, demonstrating superior zero-shot ImageNet accuracy and robustness across different data modalities.

The results underline the efficacy of using sparse models with expert layers that can be dynamically allocated to process specific types of data, suggesting a scalable approach to multimodal learning. This approach shows promise for applications requiring efficient processing of diverse data types without the need for separate specialized models for each modality.

2.3.5 Prompting Large Vision-Language Models for Compositional Reasoning

In this paper, (Ossowski et al., 2024) presents KEYCOMP, a generative method that leverages large vision-language models (VLMs) to enhance compositional reasoning through keyword-guided image descriptions. By detecting keywords from the query text and generating image descriptions focused on these keywords, the system significantly improves the alignment between visual content and text descriptions. This approach outperforms traditional embedding-based methods on the Winoground dataset by facilitating more accurate and complex reasoning.

KEYCOMP uses a large language model (LLM) like GPT-4 to analyze the generated image descriptions, enabling multi-step reasoning that addresses the challenges of matching images with their corresponding text descriptions. The paper showcases how this method can surpass embedding-based approaches, particularly in tasks that require understanding the nuanced relationships within visual scenes.

Overall, KEYCOMP marks a significant advance in multimodal reasoning, demonstrating how fine-grained control over image description generation can lead to better performance in vision-language tasks. This work suggests promising directions for future research in enhancing the descriptive capabilities of VLMs and the reasoning abilities of LLMs.

2.3.6 Linear Spaces of Meanings: Compositional Structures in Vision-Language Models

In this paper, (Mustafa et al., 2022) investigate decomposable embeddings in pre-trained vision-language models (VLMs), presenting a framework that approximates composite concepts as linear combinations of simpler vectors termed ”ideal words.” These vectors serve as foundational elements in the embedding space, enabling the linear composition of new concepts. The study primarily focuses on contextual text embeddings and their potential for compositional structures, a relatively unexplored area compared to traditional non-contextual embeddings.

The research highlights the conditional independence of decomposable embeddings, demonstrating that such structures can arise under specific probabilistic conditions. Empirical analyses using models like CLIP show that embeddings often possess decomposable qualities beneficial for classification, debiasing, and retrieval tasks in multimodal contexts. The findings reveal that straightforward linear operations on embeddings can effectively manipulate semantic meanings, offering more controlled and interpretable model interactions.

2.3.7 VL-Few: Vision Language Alignment for Multimodal Few-Shot Meta Learning

In this paper, (Ma et al., 2024a) the authors address the challenges in traditional multimodal learning which demands extensive aligned multimodal data, such as image-text pairs, by proposing a novel framework called VL-Few. This method introduces a series of alignments to enhance multimodal understanding in few-shot scenarios, significantly reducing the necessity for large datasets. The key features include modal alignment that integrates visual features into the language space, and few-shot meta learning that constructs a varied task pool to improve model generalization. Additionally, semantic, task, and generation alignments are incorporated to refine the model’s understanding of the tasks and context, as well as to improve its generation capabilities.

Experimental results confirm that VL-Few efficiently addresses multimodal few-shot problems, showing substantial improvements in handling tasks with limited data. This approach allows for efficient training of multimodal systems, making it feasible to deploy advanced AI capabilities even in resource-constrained environments.

2.3.8 An Examination of the Compositionality of Large Generative Vision-Language Models

In this paper, (Ma et al., 2024b) delves into the compositionality of Generative Vision-Language Models (GVLMs)—a category of models that combine visual and textual information. The authors recognize a crucial gap, while Large Language Models (LLMs) have achieved remarkable success, GVLMs—constructed via multimodal instruction tuning—remain relatively unexplored in terms of their ability to understand and compose visual and linguistic information. Existing evaluation metrics often prioritize syntactic correctness, but semantic compositionality is equally vital. The authors

hypothesize that generative score methods can effectively evaluate compositionality in GVLMs.

To address biases and enhance evaluation, the authors propose a novel metric: the SyntaxBias Score. Leveraging LLMs, this score quantifies the extent of syntactic bias present in GVLMs. Additionally, they introduce the Morphologically De-biased Benchmark (MODE), which includes challenging tasks to assess GVLMs' robustness against inherent inclinations toward syntactic correctness. By providing the first unbiased benchmark for GVLM compositionality, this study paves the way for future research in understanding and improving multimodal reasoning models

2.3.9 Text encoders bottleneck compositionality in contrastive vision-language models

In this paper, (Kamath et al., 2023) investigates the impact of text encoders on the compositional understanding of vision-language models. The central focus is on models like CLIP, which represent captions using a single vector. However, this compression of textual information into a single vector may lead to information loss. To address this, the authors curate a set of increasingly compositional image captions called CompPrompts. These prompts span from simple descriptions (e.g., single object) to more complex compositions (e.g., object relationships, attribute-object associations, counting, and negations). The core inquiry revolves around the ability of VLMs to effectively capture compositional factors using their text encoders.

This paper reveals that CLIP's text encoder falls short on more compositional inputs. It struggles with capturing nuances related to object relationships, attribute-object associations, counting, and negations. Interestingly, some text encoders perform significantly better than others in terms of recoverability from single-vector text representations. Further, the ability to recover captions from text-only representations predicts multi-modal matching performance. To evaluate this, the authors introduce a new evaluation benchmark called ControlledImCaps, consisting of fine-grained compositional images and captions. In summary, while text-only recoverability is necessary, it's not sufficient for modeling compositional factors in contrastive vision-language models. This work highlights the importance of addressing the bottleneck introduced by text encoders in VLMs

2.3.10 SyCoCa: Symmetrizing Contrastive Captioners with Attentive Masking for Multimodal Alignment

Multimodal alignment between vision and language is a critical research area with applications across various domains. Existing methods like Contrastive Captioners (CoCa) have made notable progress by integrating contrastive language-image pretraining (CLIP) and image captioning (IC) into a unified framework. However, a key limitation of these approaches is the lack of fine-grained multimodal alignment from both global and local perspectives. While CLIP imposes bidirectional constraints on global representations, IC only conducts unidirectional image-to-text generation on local representations, lacking constraint on local text-to-image reconstruction. This restricts the ability to achieve comprehensive understanding of images when aligned with text descriptions.

To address this gap, the authors of (Ma et al., 2024c) propose a novel framework called Symmetrizing Contrastive Captioners (SyCoCa), which introduces bidirectional interactions on images and texts across global and local representation levels. In addition to the existing ITC (image-text contrasting) and IC heads in CoCa, SyCoCa incorporates a Text-Guided Masked Image Modeling (TG-MIM) head. This enables leveraging textual cues to reconstruct contextual images and visual cues to predict textual contents, promoting bidirectional local interactions. Furthermore, the authors employ an attentive masking strategy to select effective image patches for interaction with text. Specifically, the most relevant patches semantically similar to text descriptions are chosen for the IC task to aid text generation, while the least relevant patches are selected for the TG-MIM task, aiming to leverage text to recover image content.

The proposed SyCoCa framework is evaluated through extensive experiments on five vision-language tasks: image-text retrieval, image captioning, visual question answering, and zero-shot/finetuned image classification. The results validate the effectiveness of the method, with notable improvements observed in tasks like image-text retrieval on the Flickr-30k dataset (e.g., +5.1% and +3.7% gains in mTR and mIR metrics, respectively, compared to CoCa). By enabling bidirectional global and local interactions between vision and language modalities, SyCoCa advances the field

of multimodal alignment, enhancing fine-grained understanding between images and their textual descriptions or captions.

2.4 Relevant techniques

2.4.1 Linguistic Binding in Diffusion Models

Spatial relationships between attributes and objects in the text can be aligned with the corresponding image representation. We re- fer to the approach followed in (Rassin et al., 2024). Their focus is on text-conditioned image generation, where they aim to create a strong mapping between linguistic binding of entities and modifiers in the prompt and visual binding of the corresponding elements in the generated image. While this is a text-to-image retrieval task, our scope will be focused on compositionality benchmarks formulated as image-to-text retrieval task.

2.4.2 Generating Images with Multimodal Language Models

The authors in (Koh et al., 2023) put forward a multimodal image generation model, named GILL, is designed to handle sequences of interleaved image and text inputs. This is achieved by adapting a frozen LLM to process these inputs by learning translation parameters that map image features to the text embedding space. This capability is vital for compositional reasoning as it allows the model to interpret and integrate visual information directly with textual descriptions, and use information from the image and text that come earlier in the sequence. GILL introduces special [IMG] tokens in the LLM's vocabulary to facilitate image generation. These tokens are used to bridge the gap between the LLM's text processing capabilities and the visual output required.

GILL is tested on its ability to handle visual dialogue scenarios, which test the model's capability in compositional reasoning by requiring it to understand and respond to progressively complex multimodal data. The results indicate that GILL can maintain context over extended dialogues and generate increasingly accurate images based on the evolving discussion .

2.4.3 LanguageBind

The LanguageBind framework described in (Zhu et al., 2024) extends video-language pretraining to multiple modalities by directly aligning them with the language modality, which is rich in semantic information. The primary technique used is contrastive learning, which ensures that all modalities are unified within a shared embedding space, enhancing semantic alignment across different data types. LanguageBind framework is particularly effective due to its use of a large-scale multi-modal dataset called VIDAL-10M, which includes video, infrared, depth, audio, and corresponding language data. This dataset helps validate the effectiveness of the LanguageBind framework by providing diverse modal data aligned to language

2.4.4 Cross-Modal Attention With Semantic Consistence for Image–Text Matching

The paper (Xu et al., 2020) presents a novel approach called Cross-modal Attention with Semantic Consistency (CASC) for the task of image-text matching. The authors highlight the importance of exploiting the global semantic consistency between image regions and sentence words as a complement to the local alignment between them. Previous fine-grained matching methods, such as Stacked Cross Attention Network (SCAN), rely heavily on the accuracy of detected image regions, and they may fail to capture important semantic concepts that are not easily detectable from pixel-level image data. To address this limitation, CASC incorporates a multilabel prediction branch that extracts semantic labels directly from the available text data, providing a global similarity constraint to enhance the local region-word alignment.

The proposed CASC framework consists of two main components: cross-modal attention for local alignment and multilabel prediction for global semantic consistency. The local alignment component utilizes cross-modal attention to attend to important image regions and words, calculating the local similarity between them from both image-to-text (I2T) and text-to-image (T2I) directions. The global semantic consistency component employs multilabel classification on the extracted semantic labels from the associated text, providing a global similarity constraint upon the local alignment. These two branches are integrated into a joint learning framework, leading to more accurate image-text similarity measurements on both local and global aspects. Extensive experiments on the Flickr30k and Microsoft COCO datasets demonstrate the effectiveness of CASC in preserving global semantic consistency, achieving superior image-text matching performance compared to sev-

eral state-of-the-art methods.

2.4.5 Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding

The paper (Zhang et al., 2024) proposes a novel method to improve the compositional reasoning abilities of vision-language models like CLIP. The key limitations of existing contrastive vision-language pretraining are that the negatives used are often very distinct from the positive samples, allowing the model to distinguish them based on simple object recognition rather than grasping fine-grained compositional details. To address this, the authors generate "hard negative" captions that involve changes in relationships, attributes, actions, and objects relative to the original caption. They then introduce two new loss functions - an intra-modal contrastive loss to clearly distinguish the positive and hard negative captions, and a cross-modal ranking loss with an adaptive threshold to maintain a minimum similarity gap between positive and hard negative image-caption pairs. Through extensive experiments on five compositional reasoning benchmarks, they demonstrate state-of-the-art performance improvements using their loss functions and hard negative generation strategy with CLIP and X-VLM models.

2.4.6 Coarse-to-Fine Contrastive Learning in Image-Text-Graph Space for Improved Vision-Language Compositionality

In this paper, (Singh et al., 2023) investigates the limitations of contrastively trained vision-language models in their ability to perform compositional reasoning over objects, attributes, and relations. Their findings indicate that contrastively trained VLMs have made significant strides in vision and language representation learning, yielding state-of-the-art performance on various multimodal tasks. However, recent research has underscored severe limitations in these models when it comes to handling compositional reasoning. Specifically, their ability to understand complex relationships between objects, attributes, and other linguistic constructs remains a challenge.

To address these limitations, the authors propose a novel technique that leverages scene graphs as a proxy for understanding image compositionality. Scene graphs are graph-structured semantic representations of images, capturing objects, their attributes, and relations within a scene. The authors introduce a graph decomposition and augmentation framework and a coarse-to-fine contrastive learning objective that aligns sentences of varying complexities to the same image. Additionally, they devise novel negative mining techniques in the scene graph space to enhance attribute binding and relation understanding. Through extensive experiments, their approach significantly improves attribute binding, relation understanding, systematic generalization, and productivity on multiple benchmarks. Notably, it achieves similar or better performance than CLIP on various general multimodal tasks

3 Task Setup and Data

The objective of our task is to enhance the compositional reasoning capabilities of vision-language models (VLMs). Compositional reasoning is a critical aspect of intelligence, allowing for the integration and interpretation of complex entities from simpler, constituent parts. In vision and language processing, this means understanding and generating content that accurately reflects the combination of elements within an image and their respective descriptions. Such capabilities are vital for advancing VLM applications in areas like image retrieval, text-to-image synthesis, and open-vocabulary classification.

3.1 Benchmark - SUGARCREPE

Our choice of dataset, SUGARCREPE (Hsieh et al., 2023), provides a benchmark for the compositional capability of vision-language models. This benchmark is designed to test models on their ability to handle complex, compositional data, pushing the limits of what VLMs can understand and articulate from visual inputs.

Other existing benchmarks exhibit inherent biases, which are most apparent when text-only models, with no access to the images, outperform vision language models on these benchmarks. Assessing other state-of-the-art models on SUGARCREPE revealed that their advancements were overestimated. Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) struggle with binding correct attributes to the correct objects, understanding relations between objects, generalizing systematically to unseen combinations of concepts, and to larger and more complex sentences.

Furthermore, SUGARCREPE provides a benchmark with diverse set of fine-grained hard negatives, aiming to facilitate a more faithful evaluation of vision-language models' compositionality. This dataset helps us assess our models' ability to comprehend and describe visual scenes accurately by testing their understanding of atomic concepts and relationships within the scenes, with a specific focus on attribute-binding. Figure 1 shows an example of a positive and negative caption pair, with minimal difference.

Figure 1: Sample of Image with corresponding Positive Caption and Negative Caption

We aimed to implement an approach that beats SUGARCREPE's compositionality benchmark, focusing on attribute binding.

Attribute binding is essential for producing precise and significant descriptions of visual scenes in natural language. This process demands that the model not only identify objects and their attributes in an image but also comprehend the relationships between these attributes and the objects, including their spatial arrangements. Successful attribute binding enhances the overall compositional abilities of vision-language models, enabling them to create coherent and contextually appropriate descriptions of intricate visual scenes.

This task involves two modalities: text and images. We examine the spatial relationships between objects and their attributes in the positive text and attempt to align image representation with these relationships. The aligned representations can be used to generate a single representation using fusion, which helps us train a classifier to distinguish between positive and hard-negative text (with minimal differences from the positive text) for any given image. Our aim is to assess whether the relationships expressed between the attributes and the objects in the text modality actually reflect the attribute binding depicted in the image.

3.2 Training Dataset - MS COCO

The MS COCO dataset is a large-scale resource that plays a pivotal role in advancing computer vision research. It was created by crowd-sourcing annotations from workers on Amazon Mechanical Turk. The images were chosen to represent a wide range of scenes and objects, including indoor and outdoor scenes, animals, vehicles, and everyday objects. Comprising 328,000 images, it serves as a comprehensive benchmark for various vision-language tasks. One of the key contributions of the MS-COCO dataset is its emphasis on context and relationships between objects.

In addition to object bounding boxes and segmentation masks, the dataset provides captions that describe the scene and the relationships between objects. This contextual information is crucial for developing more advanced computer vision models that can understand and reason about the semantic relationships between objects in an image. Owing to its large scale, diverse scenes, high-quality annotations, and emphasis on contextual relationships between objects, the MS-COCO dataset serves an ideal choice for training of our vision-language model.

The dataset contains annotations for:

- Object Detection: Bounding boxes and per-instance segmentation masks for 80 object categories.
- Image Captioning: Natural language descriptions of images.
- Keypoints Detection: Over 200,000 images with 250,000 person instances labeled with keypoints (17 possible keypoints, e.g., left eye, nose, right hip).
- Stuff Image Segmentation: Per-pixel segmentation masks for 91 stuff categories (e.g., grass, wall, sky).
- Panoptic Segmentation: Full scene segmentation with 80 thing categories (e.g., person, bicycle) and a subset of 91 stuff categories.

- Dense Pose: Over 39,000 images with 56,000 person instances labeled with DensePose annotations, providing detailed mappings between image pixels and 3D body templates. Notably, the annotations are publicly available only for training and validation images

3.2.1 Modality Analysis

1. No. of Objects per Image: The graph illustrated in Figure 2 represents the distribution of the number of objects across images in the dataset. Notably, the first bar (representing the range 0-10 objects) is significantly taller, indicating a higher number of images with fewer objects. This graph is essential for understanding data balance and diversity within the dataset.

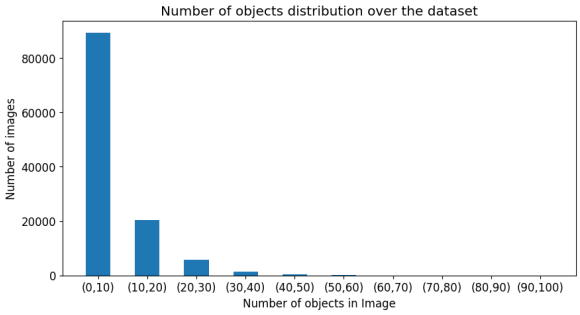


Figure 2: No. of Objects per Image

2. No. of Classes per Image: The graph illustrated in Figure 3 represents the distribution of classes (total 80) across images in the dataset in a decreasing order. Notably, person, car and chair are the top three classes present in the images across the dataset. This graph also highlights the different kinds of broad category of objects represented in the images.

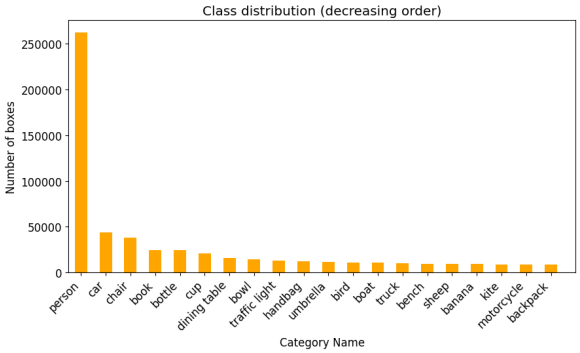


Figure 3: No. of Categories per Image

3. Average Bounding Box area per Class: The graphs illustrated in Figure 4 represent the average bounding box areas per class across images. Notably, larger objects belonging to classes such as couch, oven and bus have a higher bounding box area. Moreover, bounding boxes for smaller objects belonging to classes like sports ball, baseball bat and skis take up smaller areas, as expected. This graph highlights the expected bounding boxes annotation sizes accross the images.

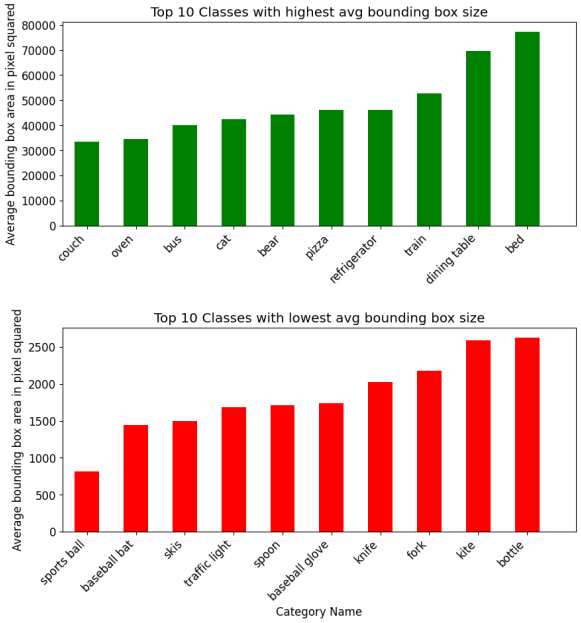


Figure 4: Average Bounding Box per Class

3.2.2 Sample Image

Figure 5 depicts a street scene with colored bounding boxes around various objects and people. Notable elements include a person sitting on a bench, a standing person (partially cropped), and a green bus in the background. The detected categories are: truck, person, person, person, person, bench, handbag, bottle, car, car, backpack. The bounding boxes align with expectations for an object detection task.

4 Baselines

For our baselines, we have 3 unimodal models, 2 simple multimodal models (with text and image modalities), and 4 competitive baselines, making use of large pretrained models. The results obtained by running these baseline on the SugarCrepe benchmark can be seen in Table 1.

The models will either generate outputs which will be compared against the positive and hard negative captions in the SugarCrepe, or will select one

Figure 5: Sample Image in MS COCO

of those captions. It's performance will be determined by how often it selects the positive caption.

4.1 Unimodal Baselines

We are experimenting with three unimodal baselines, two using the text modality, and one using an image modality.

4.1.1 Multilayer Perceptron

We developed a model employing a multi-layer perceptron (MLP), using the PyTorch library, to classify captions as either positive or hard-negative. Our training data is sourced from the GenericsKB dataset (Bhakthavatsalam et al., 2020), a comprehensive knowledge base of generic sentences.

To prepare the training dataset, we assigned labels to sentences based on their score values: sentences with $scores \geq 0.5$ were labeled as positive (1), while the rest were labeled as hard-negative (0). To ensure balanced training, we sampled an equal number of positive and hard-negative instances from the dataset, resulting in a final dataset comprising 320,000 sentences, evenly split between positive and hard-negative instances.

After acquiring the training dataset, we employed the BERT Tokenizer to tokenize each sentence, resulting in a 24-token-wide vector to represent individual sentences.

Our MLP architecture comprises an input layer, succeeded by four hidden layers, and culminating in an output layer. This model outputs a score ranging from 0 to 1, serving as a measure of the commonsense correctness of the input caption.

The key insight from running this baseline is that even though MLPs offer a straightforward approach to text classification, their limitations become evident for projects such as ours which necessitates advanced compositional reasoning. We see that both positive and hard-negative captions are syntactically and semantically correct, and thus our model is not able to correctly distinguish between the two, resulting in a chance accuracy of around 50%.

4.1.2 VisualBERT-based model

We used a custom image captioning model, built upon the foundational principles of VisualBERT (Li et al., 2019) and pre-trained on the Flickr8k dataset. This model takes in only images for its input, and employs a Vision Transformer (ViT) for extracting intricate features from images and a GPT-2 tokenizer for nuanced text processing. The ViT, functioning as the image encoder, analyzes visual content to derive meaningful representations crucial for subsequent caption generation. Meanwhile, GPT-2, rooted in the Transformer architecture, captures intricate dependencies within sequential data. The tokenizer facilitates a seamless conversion of textual information into tokens, which is important for effective model processing. In both training and inference, the tokenizer plays a pivotal role in ensuring captions are appropriately formatted and tokenized. During pre-training, the model was fine-tuned with a beam search configuration, resulting in the elevation of the quality of generated captions.

The performance inconsistency of the model across the measures could be due to the absence of targeted fine-tuning, without which the model encounters difficulties in reconciling visual and textual information, impeding its ability to grasp the intricate relationships between the two modalities.

4.1.3 Vera

Vera is a general-purpose model that estimates the plausibility of declarative statements based on commonsense knowledge (Liu et al., 2023). It only makes use of the text modality. It has been trained on 7M commonsense statements created from 19 QA datasets and two large-scale knowledge bases. Vera is selecting the caption it considers more plausible, without looking at the image. We observe that it identifies the positive captions at a rate close to chance when run against the Sugarcrepe benchmark. This would indicate that the benchmark achieves its aim of generating hard negatives which

are semantically correct and have subtle variations from the positive captions, which nonetheless generate a description different from what is in the image. On benchmarks like ARO+CREPE, language models have been able to outperform state-of-the-art vision-language models.

4.2 Simple Multimodal Baselines

We include 2 simple multimodal baselines, both making use of CNN as encoders and LSTM as decoders. One of the models also makes use of soft attention (making use of the approach detailed in Show, Attention and Tell).

4.2.1 Simple CNN-LSTM model

This model makes use of an early fusion of the image and text representations, and is implemented using the Keras library.

For the CNN encocder, we make use of the (Chollet, 2017) architecture provided by the library. This is used to extract the feature representations for the input images. We are using its default image size of 299x299, and resizing the input images are required.

For the captions associated with the images, during training, a tokenizer is used to convert the text to a sequence of integers, to be fed to the model.

The decoder is an LSTM layer of 256 units. This layer will take as its input a combined representation of the image and text representations.

For training data, we make use of 25,000 images (sampled from a total of 118K images), and their associated captions, from the COCO 2017 dataset. This curated subset was obtained by making use of coco-minitrain (Samet et al., 2020). The authors mention that the performance of a model trained on this subset is positively correlated with the performance of a model trained on the full set.

The model performs rather poorly across the measures for the SugarCrepe benchmark, indicating that it has poor compositional reasoning. There is poor alignment between the image and text modalities, leading to the model being unable to detect subtle variations in the text. we could see this inthe captions produced by this model, which were somewhat accurate, but were not detailed in capturing all relevant details in the image. This indicates that for any model we choose to implement, which improves compositional understanding of an image, we need some alignment between the image and text representations, to provide a mapping between the object-attributes in the image and the text.

4.2.2 Enhanced CNN-LSTM with Visual Attention Model

This model inspired by the (Xu et al., 2016) approach, introduces an enhanced attention mechanism to the task of image captioning, and is implemented using the PyTorch library.

For the encoder part of the model, we make use of the (Simonyan and Zisserman, 2015) architecture, pre-trained on (Deng et al., 2009), to extract feature representations from input images. Images are resized to the default input size of VGG19 (224x224 pixels) before being processed. The VGG19 encoder transforms each input image into a set of feature maps that serve as the basis for attention-driven caption generation.

For the decoder part of the model, we use an LSTM network with 256 units, designed to generate captions based on the features extracted by VGG19 and the context provided by the attention mechanism.

We implement a soft attention mechanism, as described in (Xu et al., 2016), which allows it to learn to focus on different parts of the image at different steps of the caption generation process. This approach enhances the model's ability to describe images accurately by paying selective attention to details relevant to the caption being generated.

The model is trained on the COCO 2017 dataset, using a subset of 25,000 images and their associated captions. This curated subset was obtained by making use of coco-minitrain (Samet et al., 2020). Similar to the simple CNN-LSTM model, images are resized to fit the input requirements of the VGG19 encoder, and captions are tokenized into sequences of integers. The training process involves optimizing the alignment between the visual features, attention weights, and the generated captions to improve caption quality.

While this approach did not yield any improvement in most of the measures, there was a significant improvement in accuracy. when measured against the ADD form of hard-negatives. It indicates that attention is effective in identifying new objects and attributes in the caption, which were

previously not present. The poor overall performance across all measures again indicates that the alignment between the image and text modalities is not effective. The network is not deeo enough to learn any in-depth representation of both the modalities.

4.3 Competitive Baselines

We run 4 publicly available competitive baselines. These include:

1. CLIP ViT-B/32 (Radford et al., 2021) uses a ViT-B/32 Transformer architecture as an image encoder and uses a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss. The model is finally tested on Sugarcrepe by feeding it an image and corresponding positive and hard-negative caption. The model returns a similarity score between the image and both the captions. The accuracy is calculated as a ratio between the number of times the positive caption was more similar to the image as compared to the negative caption and the total number of testing instances. We see a decline in model performance in the Swap and Add category, indicating that there's no binding between the attributes and objects in the text, since the model is not able to learn effective local representations within the image and text modalities (is not effective at linking fine-grained details). Changing relative positions of objects and attributes, and adding new attributes and objects results in a poor performance. The key insight that we can conclude based on the results is that keeping a track of which attributes are bound to which objects, and the relative positions of those objects and attributes in the caption is important for better performance.

2. RN50 (He et al., 2015) is a Residual Network consisting of 50 layers. It consists of residual blocks with skip connections that allow gradients to flow through the network more effectively during training. The key insight from running this baseline is that RN50 is very effective at identifying existing objects in the image, and mapping them to the captions i.e. if an existing object gets replaced in the caption, the model identifies that, and selects the positive caption. However, it struggles if new objects are added, or if the existing objects are swapped in the caption text. There is also degraded performance in the benchmark when it comes any changes made in the relations (for example, replacing "running" with "walking").

3. RoBERTa-ViT-B-32 integrates RoBERTa's (Liu et al., 2019) NLP capabilities with the Vision Transformer's (ViT-B-32) (Dosovitskiy et al., 2021) image analysis. It is trained on the LAION-2B dataset and scaled to 12 billion parameters. It is designed to perform well in both understanding complex textual content and performing intricate visual analyses. The key insight from running this baseline is that RoBERTa-Vit-B is very effective in incorporating new objects into captions and adjusting to changes in object relationships within the text, that is, it shows a significant leap in adaptability and understanding. However, the captions generated followed a common pattern that was seen during the training. In some cases, it lacked creativity or failed to capture more abstract or nuanced aspects of the images.

4. CLIP Enhance-FineGrained, based on (Zhang et al., 2024) uses two loss functions to enhance the compositional understanding ability for any contrastive vision-language models loss like CLIP. This method significantly improves compositional reasoning in Vision-Language Models by refining the image-text contrastive learning framework through intra-modal contrast and cross-modal rank objectives. Intra-modal contrast enhances image-text alignment within the same modality, while cross-modal rank improves the model's understanding by ranking correct pairs higher than incorrect ones across modalities. This enhancement results in notable performance gains on challenging benchmarks like SugarCrepe, where the model excels in tasks such as adding objects, adding attributes, replacing objects, and replacing attributes. The substantial accuracy improvements demonstrate enhanced vision-linguistic understanding and reasoning capabilities. These results underscore the efficacy of the proposed approach in advancing compositional reasoning in Vision-Language Mod-

els, particularly in tasks requiring nuanced semantic variations and fine-grained image-text alignment.

5 Proposed Model

5.1 Overall Model Structure

In this project, our aim is to implement a visual-language model with improved compositional reasoning. The problem we aim to solve is to give our model the capability to identify slight variations in text associated with an image. Although these slight variations result in syntactically and semantically correct texts, the corresponding images would be fundamentally different. Current visual-language models are not able to reason well enough to identify these slight but significant variations in captions.

In order to overcome these shortcomings, we explore three possible model structures -

- **Basic Contrastive Learning Model :** The first model structure involves a straightforward application of contrastive learning to enhance attribute binding in vision-language tasks. In this setup, an image is fed into the image encoder (detailed in Section 5.2.1) while the text is processed through a text encoder. The embeddings produced by the text encoder (detailed in Section 5.2.2) undergo a linear projection to transform them into a space that is comparable with the image embeddings. A similarity score between the image and text embeddings is then calculated. The primary objective is to minimize the distance between the correctly paired text and image embeddings while maximizing the distance between mismatched pairs, using a contrastive loss function. This model focuses on basic multimodal alignment but lacks deeper integration of cross-modal interactions, potentially limiting its ability to handle more complex compositional reasoning. This approach is illustrated in Figure 6

- **Enhanced Cross-Modal Attention Model :** The second model extends the basic framework, explained above, by incorporating a cross-modal attention mechanism, which allows for a more dynamic interaction between the text and image modalities. After encoding and projecting the text embeddings as in the

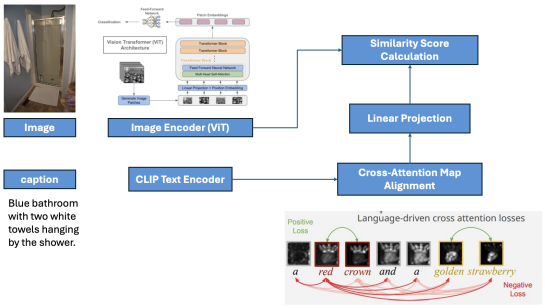


Figure 6: Basic Contrastive Learning Model

model above, a cross-modal attention layer is introduced to refine the alignment between the text and image representations. This attention mechanism enables the model to focus on specific parts of the text and image that are most relevant for matching, thereby encouraging alignment between the two modalities. The contrastive loss remains the primary learning objective, ensuring that the model learns to correctly match corresponding text-image pairs while distinguishing non-matching pairs. This approach aims to help us learn more correct multimodal representations by facilitating a more granular and context-aware interaction between the two modalities.This approach is illustrated in Figure 7

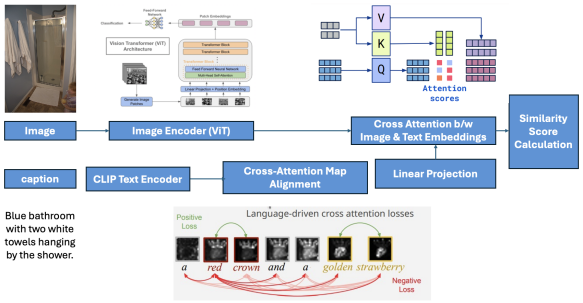


Figure 7: Enhanced Cross-Modal Attention Model

- **Fusion and Shared Representation Learning Model :** Building upon the previous models, the third model is aimed to learn a unified representation of text and image inputs. After processing through respective encoders and a cross-modal attention mechanism, both the image and text embeddings are concatenated. This combined representation is then passed through a multi-layer perceptron (MLP) with five hidden layers, aiming to fuse the modalities into a cohesive shared embedding space.

The primary learning objective in this model is cross-entropy loss, applied after a softmax layer to predict the correct caption among multiple candidates. In addition to the primary loss function, we employ contrastive loss as an auxiliary objective, applied to the image and text embeddings obtained from the cross-modal attention module. This model not only maintains the benefits of cross-modal attention from the second model structure proposal above but also leverages deep fusion of the embeddings, potentially leading to superior performance in tasks requiring nuanced understanding of complex object-attribute relationships in both textual and visual domains. This approach is illustrated in Figure 8

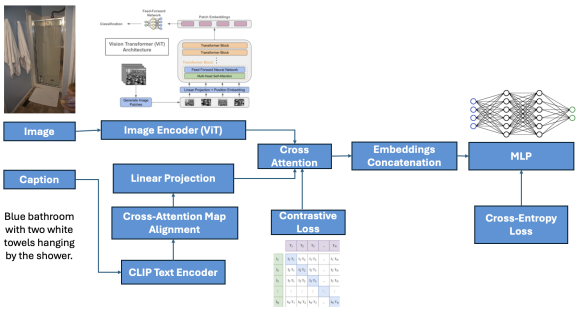


Figure 8: Fusion and Shared Representation Learning Model

5.2 Encoders

5.2.1 Image Encoder

In the development of our model, we carefully considered the choice of image encoder to best support our objectives of enhancing compositional reasoning and attribute binding.

In our evaluation of image encoders, we compared the Vision Transformer (ViT) with ResNet-50. ViT excels in analyzing complex visual relationships due to its patch-based approach and attention mechanisms, which allow for a nuanced understanding of spatial hierarchies within images. This capability enables it to capture both minute details and broader global contexts effectively. On the other hand, ResNet-50 is proficient in hierarchical feature extraction but tends to focus more on local features, which may not comprehensively capture the global context of an image. Given the needs of our project, which emphasize global contextual understanding and detailed component integration, ViT emerged as the better choice due to its robust performance in handling complex image compositions.

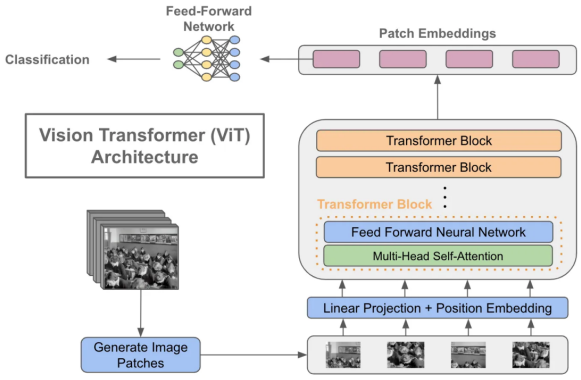


Figure 9: ViT Architecture

As shown in Figure 9, The Vision Transformer (ViT) processes images by first dividing them into fixed-size patches, which are then flattened and transformed into linear embeddings, similar to words in a sentence. Positional embeddings are added to maintain the sequence's spatial context, essential for tasks that transformers typically handle.

ViT's core consists of several transformer blocks, each equipped with multi-headed self-attention and multi-layer perceptrons (MLP), with layer normalization applied before each block and residual connections afterward. The self-attention mechanism prioritizes different image patches based on their relevance, allowing ViT to maintain a comprehensive understanding of the entire image.

This structure enables ViT to excel in tasks that require nuanced understanding of image relationships and attributes, making it highly suitable for advanced compositional reasoning and generating accurate, contextually relevant image descriptions.

5.2.2 Text Encoder

For our model, we utilize a text encoder designed to enhance the alignment of visual attributes with their corresponding linguistic descriptors in text-conditioned image generation models. This encoder employs a novel loss function developed by (Rassin et al., 2024) that targets the cross-attention maps between entities and their modifiers, optimizing these during inference without the need for retraining the model.

The text encoder operates by first syntactically analyzing the input prompt to identify entity-nouns

and their associated modifiers. This analysis is conducted using a transformer-based dependency parser which dissects the sentence structure, allowing the system to determine the syntactic relationships essential for accurate attribute binding.

Once entities and modifiers are identified, their relationships are controlled through targeted adjustments to the cross-attention maps which essentially links specific terms in the caption to corresponding areas in the image.

The custom loss function designed for this encoder is pivotal in achieving the desired attribute binding. As shown in Figure 8, it includes two main components:

Positive Loss: This loss minimizes the distance between the attention maps of modifiers and their corresponding nouns, promoting a high overlap where it is grammatically appropriate. Essentially, it encourages the model to align the visual representation of a modifier closely with its target noun.

Negative Loss: Conversely, this loss increases the distance between the attention maps of related words and those of other unrelated parts of the text. This aspect of the loss function helps to prevent the modifiers from influencing unrelated nouns or being affected by other modifiers inappropriately.

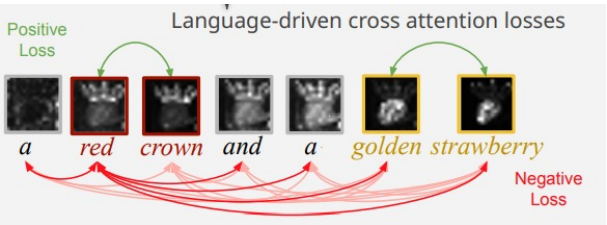


Figure 10: Language Driven Cross Attention Loss

The combination of these losses ensures the detailed and nuanced understanding of the relationships between different components within an image are captured.

5.3 Loss functions

5.3.1 Language Driven Cross Attention Loss (L1)

After getting the text embedding from our text encoder, we align the attention maps of the entities and modifiers in the text using the process outlined in (Rassin et al., 2024). Given an input text of N tokens, for which k noun-modifier sets $\{S_1, S_2, \ldots, S_k\}$ are extracted. Let $P(S_i)$ represent all pairs (m, n) of tokens between the noun root n and its modifier descendants m in the i-th set S_i. For illustration, the set of "A black striped dog" contains two pairs ("black", "dog") and ("striped", "dog"). Next, denote by $\{A_1, A_2, \ldots, A_N\}$ the attention maps of all N tokens in the prompt, and denote by $\text{dist}(A_m, A_n)$ a measure of distance (lack of overlap) between attention maps A_m and A_n.

The first loss aims to minimize that distance (maximize the overlap) over all pairs of modifiers and their corresponding entity-nouns (m, n),

$$L_{\text{pos}}(A, S) = \sum_{i=1}^{k} \sum_{(m,n)\in P(S_i)} \text{dist}(A_m, A_n) \quad (1)$$

$$\mathrm{L_{neg}} = -\sum_{i=1}^{k} \frac{1}{|U(S_i)|} \sum_{(m,n)\in P(S_i)} \sum_{u\in U(S_i)} \frac{1}{2} \left(\text{dist}(A_m, A_u) + \text{dist}(A_u, A_n)\right) \quad (2)$$

The final loss, L_1 combines the two loss terms:

$$L = L_{\text{pos}} + L_{\text{neg}} \quad (3)$$

5.3.2 Contrastive Loss

The loss function has two terms for each pair, one for the image-to-text direction and one for the text-to-image direction, effectively doubling the contribution of each positive pair and promoting symmetrical co-learning between the modalities.

$$\mathrm{L_{CLIP}} = -\sum_{i=1}^{N} \left[\log \frac{\exp(\text{sim}(v_i, w_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, w_j)/\tau)} + \log \frac{\exp(\text{sim}(v_i, w_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_j, w_i)/\tau)} \right] \quad (4)$$

where v_i and w_i represent the normalized embeddings of the i-th image and text, respectively. The function $\text{sim}(v, w)$ computes the cosine similarity between two vectors, and τ denotes a temperature parameter that scales the similarity scores.

5.4 Changes to training data

Our model is being trained on MS COCO's 2017 training set, which has 118,000 images. However, we decided to take a representative subset of 25,000 images, due to time and computational constraints. This subset was sampled using coco-minitrain (Samet et al., 2020), which has been shown to provide a training subset that provides results comparable to training on the entire dataset. We have not added any transformations/augmentations to the data.

5.5 Hyperparameters and their effects

Choice of Pooling for Text Encoder In the development of our text encoder, we explored various hyperparameters to optimize the processing of textual data into a format suitable for our task. We particularly focused on the pooling methods employed to condense text data.

The initial phase in our text encoder involves extracting features from each segment of text, using the CLIP model to encode these segments. Each segment is tokenized, resulting in a variable number of tokens per segment. To standardize the input for processing, we pad these tokens to match the length of the longest segment, thereby ensuring that each input tensor has a uniform shape. The resulting 3D vector representation from CLIP has dimensions: the number of segments processed, the number of tokens in the longest segment (after padding), and the number of features (embedding dimension) per token.

To transform this 3D vector representation into a 1D vector suitable for our model, we employed pooling strategies. The necessity of pooling arises from the need to reduce the second dimension, which represents the number of tokens per segment. Specifically, we experimented with two types of pooling:

Max Pooling: Applied across the token dimension (second dimension) of each segment, max pooling selects the maximum value for each feature embedding across the tokens. This method effectively captures the most salient features within each segment, reducing the data's complexity while preserving crucial information.

Mean Pooling: After max pooling, we applied mean pooling across the segments dimension (first dimension). This step averages the max-pooled features from each segment, collapsing the data into a single vector that represents the entire input text. Mean pooling helps in condensing the information from all segments, providing a comprehensive representation of the overall textual input

After running ablations by changing these hyperpramaters, We chose to go with max pooling first followed by mean pooling. By employing max pooling, our model emphasizes the extraction of the most prominent features within each token array. It the model to focus on the most critical features that define each segment's semantic properties. Mean pooling averages out the contribution of each segment, mitigating the risk of overemphasizing features from any single segment and providing a holistic view of the text.

Following max pooling with mean pooling across the segments provides a balanced method to condense these salient features into a single, comprehensive feature vector. This combination ensures that the encoder captures a broad representation of the entire text, integrating essential features from each segment.

Base Image Encoder We experimented with two image encoders - CLIP ViT-B/16 and ViT-B/32. Our final model is making use of CLIP ViT-B/32. The decision was based on our analysis of the image-text similarity score for each of the, which is discussed in detail in Section 7.1.5. ViT-B/32 had the higher score, meaning the image imbeddings and the corresponding text embeddings were better aligned.

Contrastive Losses We experimented with the auxiliary loss for the module that performs cross-attention between the image and text embeddings. We tried Contrastive Loss and Margin Loss, since contrastive loss can give results with low confidence, and we wanted to use margin loss to ensure that dissimilar items are not just far apart, but specifically farther than the dissimilarity within similar pairs by a set margin. However we decided to go with Constrastive Loss for our final model, because the use of margin loss resulted in a slightly decreased accuracy, contrary to our expectation, though it did give those results with greater confidence.

6 Results

We measure the performance of our selected baselines and proposed models against the benchmark provided by SUGARCREPE, to check their compositional reasoning. The results are illustrated in the Table 1 below.

Our final proposed architecture applies a softmax activation thus selecting one of the captions as positive. For our other explored architectures and the baseline models using both image and text modalities, we check the cosine similarity of the generated captions against the positive and hard negative captions provided for that image by SUGARCREPE. If the similarity is higher for the positive caption, it

Table 1: Results Table (Accuracy %)

Methods	REPLACE			SWAP		ADD	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute
Multilayer Perceptron (Text only)	49.16	49.25	50.07	46.09	49.63	50.27	48.79
VisualBERT-based model (Image only)	60.35	52.53	56.04	60.00	65.46	43.74	48.55
Vera (Liu et al., 2023)	51.45	50.89	51.06	50.20	51.20	51.06	51.44
Simple CNN-LSTM model	48.79	46.94	54.59	47.75	45.19	22.59	39.65
CNN-LSTM with Visual Attention (Xu et al., 2016)	49.69	18.65	49.93	35.99	18.77	63.92	83.53
CLIP ViT-B/32 (Radford et al., 2021)	90.74	80.33	69.42	61.22	64.11	77.01	69.51
RN50 (He et al., 2015)	91.76	80.58	69.91	62.04	68.47	74.49	69.80
RoBERTa-ViT-B-32 (Liu et al., 2019)	92.85	84.89	72.40	62.85	71.02	87.34	79.91
CLIP Enhance-FineGrained (Zhang et al., 2024)	93.09	88.83	79.01	73.06	77.02	92.38	93.35
Basic Contrastive Learning Model (Proposed v1)	71.52	66.30	62.81	55.79	56.22	42.65	45.05
Enhanced Cross Modal Attention Model (Proposed v2)	83.49	75.78	64.27	54.26	52.42	66.09	57.40
Fusion and Shared Representation Learning Model (Final)	85.66	72.37	70.04	56.51	58.43	79.17	68.70

will be considered an accurate result. For text-only models, our objective is to assess their ability to classify captions as either positive or hard-negative based on commonsense reasoning.

We will be evaluating this across 3 forms of hard negatives: Replace, Swap and Add.

REPLACE Accuracy The hard negative captions will have objects, attributes and relations (OARs) from the positive caption replaced with different values. This measure will check the model's ability to be able to just identify the OARs correctly (since replacing these in the captions will remove the original entirely). We see that all our proposed architectures perform higher than the unimodal and simple multimodal baselines, however they were not able to beat the performance of competitive baselines. We see a similar degradation in performance amongst our model proposals, when compared with the baseline implementations, for replace attribute and replace relation categories. Our final proposed model is able to beat CLIP ViT-B/32 and RN50 competitive baselines in the replace relation category, however the difference in performance is almost negligible.

SWAP Accuracy The hard negative captions will have objects and attributes (OAs) from the positive caption swapped with other OAs from the caption. Swapping the relations will not result in semantically coherent sentences in a lot of cases, so this will not be done here. This measure will provide a more robust estimation of the model's ability to recognize the object-attribute bindings in the image, as compared to replacing the values with new ones, as all the objects and attributes originally present in the image, will still be present in the hard negative caption, albeit presented in a different order/context. Similar to the baselines our proposed models show a degraded performance when it comes to these SWAP measures.

ADD Accuracy The hard negative captions will have new objects and attributes (OAs) in addition to the existing OAs in the positive caption. Again, adding relations might not result in sensible captions, so it will not be done here. It will allow us to determine if the model can identify the presence of new (and incorrect) OAs, even if the existing ones in the image are still present in the caption. The baseline show a good performance when it comes to identifying added objects, and a somewhat degraded performance in identifying added attributes. Two of our proposed approaches showed poor performance when compared with the baseline, however our final proposed approach showed improvement over CLIP ViT-B/32 and RN50 competitive baselines in the add object category, and an almost similar performance for the add attribute category.

Our final proposed model was able to beat the ViT-B/32 and RN50 competitive baselines in the add object category and showed a comparable performance to these two baselines among other categories. Since, in this approach, we learn a shared embedding between the two modalities, the shared embedding would vary significantly when we introduce a new object in the caption. However, for other categories, we see that the hard-negatives do not have such a significant difference. For example, in the replace object category, the positive caption may have the object "woman," and the

corresponding hard-negative caption may have the object "man." Though these objects are different, they may be close to each other in the embedding space. Thus, applying cross-attention, and the subsequent generation of shared embeddings has no significant improvement over the baseline models' performance. We would need to explore different architectures and loss functions to better align the image and text modalities.

7 Analysis

For our analysis, we will first evaluate the performance of our chosen baseline models using intrinsic metrics to measure the performance of the models's components on fundamental tasks. We will then proceed to analyze some failure cases for our baseline models, and identify the reasons for those failures.

7.1 Intrinsic Metrics

These metrics are meant to evaluate different aspects of the baseline models' performance, independent of an external application or context, and will be used to get insights into the models' strengths or weaknesses.

Not all metrics will apply to all models, and the corresponding slots in the table are left blank accordingly.

7.1.1 Object Detection Score

This intrinsic metric will be used to evaluate the efficacy of multimodal models and our unimodal VisialBert-based model in detecting and classifying objects in images. We will be measuring the **performance of the visual encoders** for this task, since accurate object detection forms the bedrock upon which which coherent and contextually relevant image captions are generated.

We have calculated the score in the following manner - we take the objects detected by our models from within an image, concatenates these identified objects into a single string, and then perform a comparison with a similarly concatenated string of object types listed in the ground truth. For the comparison, we obtained a similarity score, using Spacy's [1] *en_core_web_lg* vectors, which will help us capture the semantic similarity when comparing object types, rather than looking for an exact match. For instance, we do not want "person" and "man" to be considered as entirely incorrect.

[1] https://spacy.io/

We have run the baseline models against a subset of the MS COCO (Lin et al., 2015) validation set. The subset consists of those images which have 10 or less objects listed in the ground truth, and we based this on the following factors:

- There are over 3800 images containing with 10 objects or less listed in the COCO validation annotations, out of a total of 5000. This makes our choice a very representative sample.
- The models were trained on a subset of 25000 images from the COCO training set (obtained through coco-minitrain (**?**)). While this is a representative set, it does have a bias towards images containing 10 objects or less.
- Some of the models being evaluated are rather basic, and their performance is expected to be poor when it comes to detecting a large number of objects (for instance, 30 objects).

The scores can be seen in Table 2. We have also plotted graphs showing the variation in the object detection scores for different numbers of objects in the ground truth. The aim was to explore if the performance changes as the number of objects that need to be detected increase.

Simple CNN-LSTM This was a simple multimodal implementation, consisting of a CNN encoder (visual backbone) and LSTM decoder. The visual backbone is being evaluated here, and it is making use of the Xception architecture (Chollet, 2017), obtained from the Keras library. This has been pre-trained on the ImageNet (Deng et al., 2009) dataset, and has gone through some finetuning on the MS COCO dataset for our purposes. The score obtained by this model is significantly lower than obtained in the original paper (where it was tested on the ImageNet dataset). The following are some insights we gain from this performance:

- The Xception model should have gone through additional finetuning on the COCO dataset. While we should expect its performance to generalize, there is bound to be a greater difference in the predicted and actual object labels from different datasets without adequate finetuning. This would be even more necessary for such a simple encoder-decoder architecture. Additionally, the features learned from ImageNet may not align

Table 2: Intrinsic Metrics

Methods	Object Detection	Lexical Density	Mean	Variance	Image-Text Similarity
Multilayer Perceptron (Text only)	-	-	0.352*	$2.4e-2$*	-
VisualBERT-based model (Image only)	0.473	-	0.501	$2.3e-4$	-
Vera (Liu et al., 2023) (Text only)	-		0.501	$1.7e-3$	-
Simple CNN-LSTM model	0.462	0.492	0.498	$2.7e-4$	-
CNN-LSTM with Visual Attention (Xu et al., 2016)	0.445	0.469	0.502	$1.5e-5$	-
CLIP ViT-B/32 (Radford et al., 2021)	-	-	0.700	$6.3e-2$	0.330
RN50 (He et al., 2015)	-	-	0.504	$3.8e-5$	-
RoBERTa-ViT-B-32 (Liu et al., 2019)	-	-	0.507	$7.9e-5$	-
CLIP Enhance-FineGrained (Zhang et al., 2024)	-	-	0.505	$2.7e-5$	-
Fusion and Shared Representation Learning Model (ViT-B/16)	-	-	-	-	0.323
Fusion and Shared Representation Learning Model (ViT-B/32)	-	-	0.610	$2e-2$	0.330

well with those necessary for detecting and recognizing objects in the diverse contexts presented in COCO.

- The Xception model requires the images to be resized to the dimension 299x299. This downsampling leads to a loss in information as compared to the original image, which in turn gives a degraded performance for object detection. The difference in image quality can be seen in Figures 11 and 12.

- The Xception model is designed primarily for image classification, predicting a single label for the whole image. Adapting it to object detection involves not only classifying but also precisely locating multiple objects within an image, a task for which it wasn't specifically optimized, even though it can be used for this.

- We can see from Figure 14 that the object detection score varies from 0.44 to 0.48, which is a quite a narrow range. This would be because the captions generated by the model as rather generic. For instance, "person", "animal", and "fruit", instead of "girl", "bear", and "banana". While these have some similarity, the generic nature will result in a generally lower score. There are also some interesting instances where the model gives a very specific category. For example, for the image in Figure 12, one of the predicted object categories was "Granny-Smith", which is a type of green apple, while the ground truth Figure 11 has simply listed "apple" as its category. We also see from the graph that the score generally drops as the number of objects in the ground truth increases from 1 to 4, though the decrease is not significant. The reason for the decrease is immediately apparent because of the increased complexity of the images. However, we also observe that an increase when the number of objects goes above 4. This is likely because of object types being repeated in the COCO annotations, combined with our concatenating of the object types in a single string for comparison, resulting in the slightly increased similarity score.

- There are some cases, as in Figures 11 and 12 where the probability of the first object classification by the model is very high, while the probability of the remaining predicted objects is extremely low. As we can see in the example, this is a correct prediction by the model, but included those values with low probability has had an adverse impact on our similarity score.

CNN-LSTM with Visual Attention This architecture used in this model is the one proposed by (Xu et al., 2016), and in our implementation, makes use of VGG-19 (Simonyan and Zisserman, 2015) as its visual backbone (encoder). We see a slightly reduced score for this model (see 2), which is expected as the Xception architecture beat VGG-19 when tested on the ImageNet dataset. The plot of the similarity scores against the number of objects listed in the ground truth can be seen in Figure 15. While this model makes use of attention mechanisms, it is being used in the decoder, and hence the insights we gain from the score and graph are similar to what we saw in the previous model (Sim-

ple CNN-LSTM).

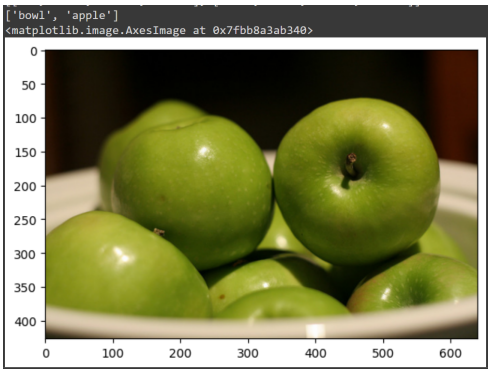


Figure 11: Original Image

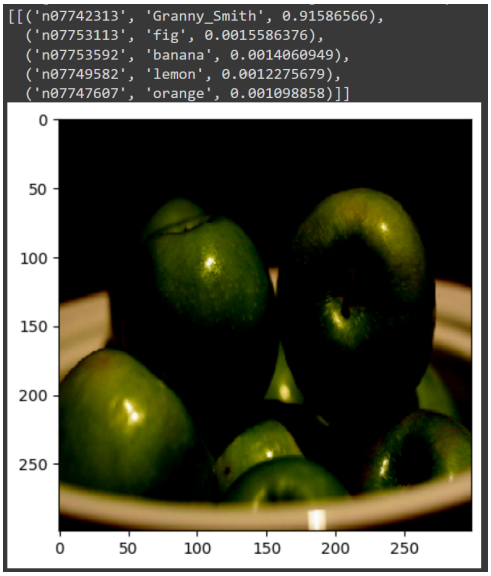
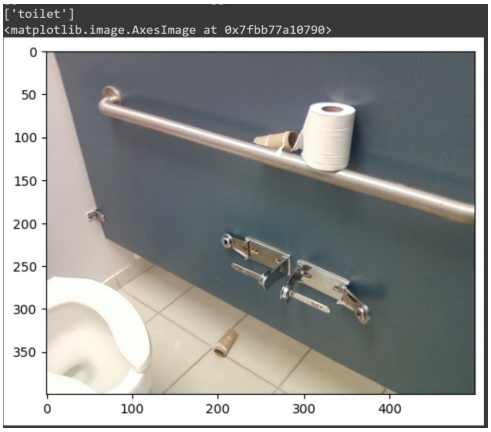


Figure 12: Resized image

VisualBERT-based model This was a custom image-captioning model built upon the foundations of VisualBert and pretrained on the Flickr8 Dataset. The finetuned version of the visual backbone is being evaluated here. The visual backbone is a Vision Transformer (ViT) i.e. a transformer encoder model (BERT-like) pretrained on a large collection of images in a supervised fashion, namely ImageNet-21k, at a resolution of 224x224 pixels. Further, the model was fine-tuned on ImageNet (also referred to as ILSVRC2012), a dataset comprising 1 million images and 1,000 classes, at a higher resolution of 384x384.

The evaluation of the model based on the object detection task reveal a consistent performance, with an average similarity score of approximately 0.476 and a narrow range from 0.45 to 0.49 as illustrated in Figure 16. While the model demonstrates stability in detecting and classifying objects across various images, the limited range of similarity scores suggests potential challenges in accurately discerning between objects in certain contexts. This observation indicates the need for ongoing improvements to enhance the model's precision and reliability for object detection tasks.

Figure 13: COCO image used for Caption Generated



This model, originally designed for image classification is bound to face challenges in object detection tasks due to its division of images into fixed-size patches. This fixed-size patch approach limits the model's ability to accurately localize objects within images, as objects spanning multiple patches may not be effectively captured. Object localization in detection tasks requires understanding spatial relationships within objects, which the patch-based processing of ViT may struggle to capture. The lack of detailed spatial context and object boundaries within patches could lead to lower similarity scores in object detection, highlighting the need for modifications to enhance ViT's performance in tasks requiring precise object localization.

7.1.2 Lexical Density

Lexical density is a measure of the complexity of a text, quantifying the proportion of content words — nouns, verbs, adjectives, and adverbs — relative to the total number of words. In image captioning, it reflects the caption's informativeness and linguistic richness, revealing the model's capability to emulate human-like descriptive language.

In examining lexical density distribution of ground truth captions alongside those generated by a simple CNN-LSTM model and a CNN-LSTM model with visual attention, several insights can be gleaned about the efficacy and linguistic characteristics of the models in comparison to human

Figure 14: Object Detection by Simple CNN-LSTM Model

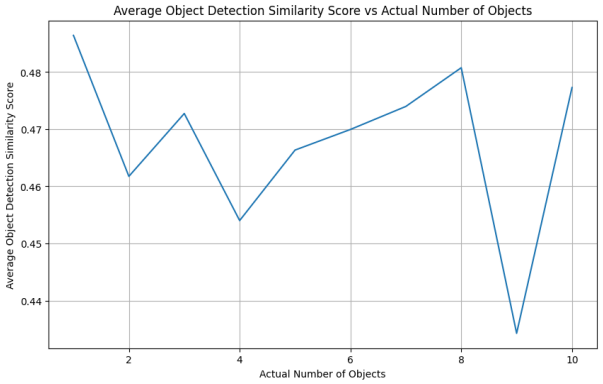


Figure 15: Object Detection by CNN-LSTM with Visual Attention

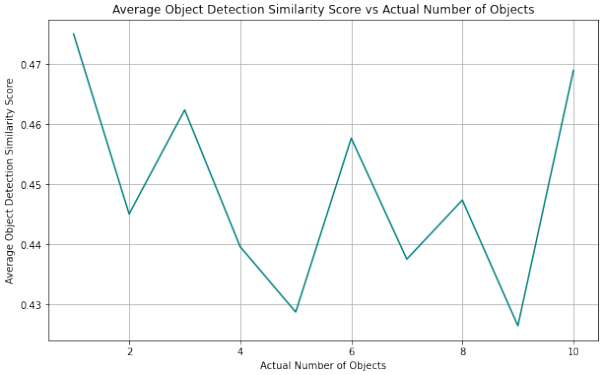
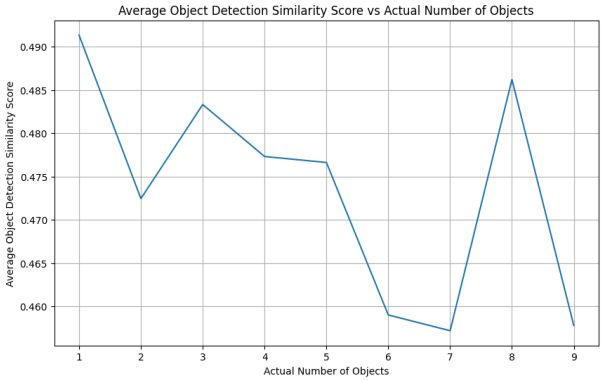


Figure 16: Object Detection by Vision Transformer



standards.

As per Figure 17, for the ground truth captions, the distribution reveals a balanced variance with a symmetrical bell curve, peaking around a lexical density score of 0.5. This suggests that human captions tend to have a moderate and consistent level of lexical richness, without tending towards either extreme verbosity or oversimplification. The smoothness of the distribution also implies a diverse usage of language across different samples.

As illustrated in Table 2, the lexical density scores for the simple CNN-LSTM model are higher than those of the CNN-LSTM model with attention. However, a higher average lexical density score does not necessarily equate to a superior model performance. Therefore, we will now analyze the distributions to ascertain the implications of lexical density variations on the quality of the generated captions.

In analyzing the performance of our simple CNN-LSTM model for image captioning tasks, from Figure 19 it is observed that the model consistently generates captions with a moderate lexical density. This is evidenced by the prominent peak at the 0.5 lexical density score seen in the histogram provided. Such a distribution suggests a certain level of consistency in the model's captioning, with a recurring use of a comparable number of content words across different images. However, this consistency in lexical density does not extend to the breadth of linguistic expression. The model's captions lack the variability characteristic of human-generated text, as indicated by the narrow range of lexical density scores. The model's current architectural configuration seems to limit its ability to produce a richer vocabulary and varied sentence structures, which are essential for capturing the nuanced descriptions often found in human captions.

Using an example caption generated by this model —"start of the kitchen with white walls and white walls and white walls"— for the image shown in Figure 13, illustrates a tendency towards verbosity that does not contribute to the informative content of the caption. Such repetitive verbosity points to a shortfall in the model's language generation capabilities, emphasizing a single attribute unnecessarily without offering a comprehensive description of the image.

The CNN-LSTM model with visual attention, however, shows improvement over the simple

model, with a distribution that attempts to approximate the ground truth pattern, peaking closer to the lexical density score of the ground truth captions. Upon analyzing the lexical density histogram for our CNN-LSTM model with attention as given in Figure 18, we observe a more variable distribution than that of the simple CNN-LSTM model. The distribution is broader with several peaks, indicating the model's capacity to produce captions with varying lexical densities. This variation suggests that the attention mechanism in the model allows for a more nuanced understanding of images, which translates into a diversity of caption complexity. However, despite this variability, the distribution lacks the smooth, bell-shaped curve seen in the ground truth data, which would indicate a well-balanced use of language across captions. The multiple peaks in the model's histogram may point to inconsistencies in caption generation, where the model fluctuates between different levels of detail and complexity, potentially producing verbose captions in certain instances. The provided example caption, with its excessive repetition, further exemplifies the model's inclination towards verbosity. It highlights a need to refine the model's ability to generate concise and contextually relevant descriptions. While the attention mechanism enables the model to focus on various aspects of the image, the current training may not adequately penalize the generation of repetitive or superfluous content.

7.1.3 Similarity Score Distribution

We plot similarity score values for those models which are using similarity scores to select between positive and hard-negative captions. We believe this offers insight into the inherent ability of our models to differentiate between positive and hard-negative captions based on their similarity scores. It serves as a means to evaluate the discriminative power of the model in distinguishing between nuanced categories, thereby assessing its ability to capture subtle differences in semantic meaning. This metric is particularly essential for tasks aimed at improving model compositionality, where the accurate representation of complex relationships and semantic nuances is crucial. Here, we are not concerned with the accuracy of the classifications, but rather it's "confidence".

An important thing to note here is that we have used the terms **Category A** and **Category B** to represent the positive and hard-negative captions. A well-separated distribution with minimal overlap indicates that the model effectively captures the differences between categories, while a more blended distribution suggests potential challenges in discrimination.

Figure 17: Lexical density of Ground Truth Captions

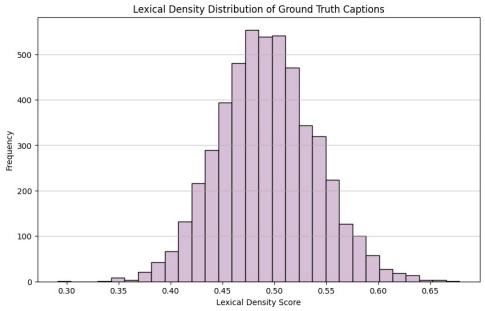


Figure 18: Lexical density of CNN-LSTM with Visual Attention

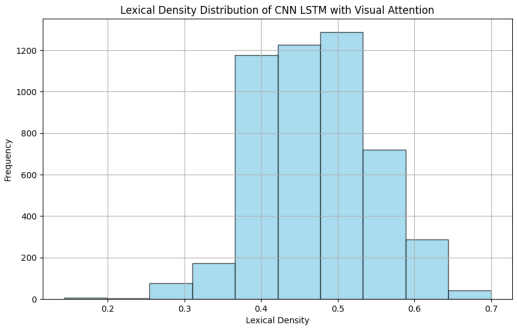
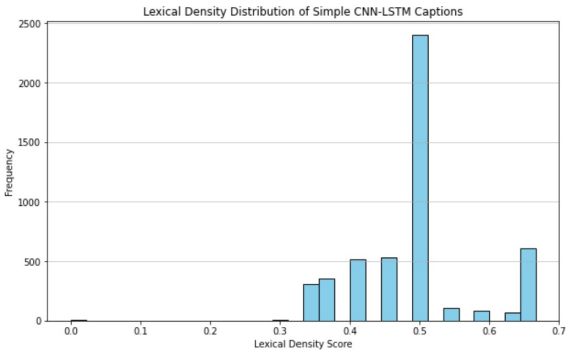


Figure 19: Lexical density of CNN-LSTM



For models - RN50 Figure 20, RoBERTa-ViT-B-32 Figure 21, Simple CNN-LSTM Figure 22 and CNN-LSTM with Visual Attention Figure 23 model we found that the data points in the scatter plot were mostly distributed along the **Boundary of Uncertainty**. From this we can draw the insight that the similarity scores generate by using both Category A and Category B are closer to each other, and there's only a marginal difference that helps us to distinguish whether an output caption is more

similar to Category A or to Category B. These models are not able to categorize with a high degree of confidence.

For Vera Figure 24, we can see that it is able to categorize the instances with greater confidence. However, we need to keep in mind that this is a language-only model, which assigns plausibility scores to texts, and does not use image features at all. Also, the captions provided are from SugarCrepe (Hsieh et al., 2023), and have very subtle differences between the different categories. The text-only unimodal baselines in our analysis are present only to highlight the importance of avoiding blatant discrepancies between the positive and hard-negative captions. If such discrepancies are present, the language-only models have been shown to give a performance equal to or better than the vision-language models (Hsieh et al., 2023).

7.1.4 Probability Distribution of Selecting Category A

For our analysis, we will focus on the probabilities assigned by the models for Category A. Using this intrinsic metric, we aim to glean insights into model confidence along with its predictive confidence.

Two of our models - Multilayer Perceptron (see Figure 25) and CLIP (see Figure 31) were directly returning probability values for whether Category A is correct or Category B. However, other models were returning similarity scores. For converting these similarity scores to probability values, we apply a softmax function over them. The value that we get for Category A is then treated as the probability of choosing Category A as the correct caption.

We also find the mean and variance values for these distributions (see Table 2. These help us understand how are model is performing in average and how the confidence of the model varies over the dataset.

We see that for most of our baseline models - VisualBERT (see Figure 28), Vera (see Figure 29), Simple CNN-LSTM (see Figure 30), RN50 (see Figure 32), RoBERTA (see Figure 33) and CNN-LSTM with Visual Attention - have high uncertainty with mean of the distribution at or around 0.5. We correlate this with probability distribution graphs and see that the distributions also peak around the value of 0.5, indicating that the models are not able to discriminate as strongly between the two categories.

Figure 20: Similarity Scores for RN50

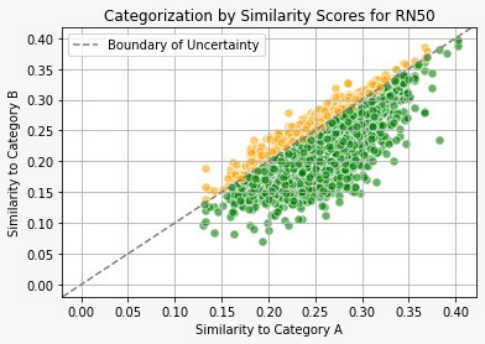


Figure 21: Similarity Scores for RoBERTa-ViT-B-32

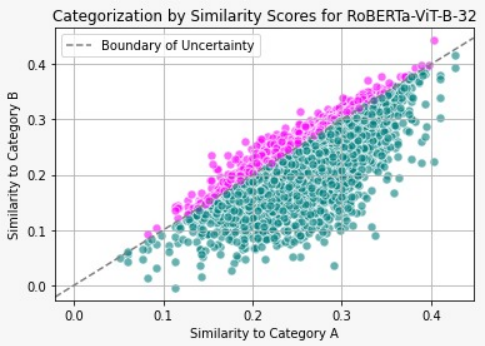


Figure 22: Similarity Scores for Simple CNN-LSTM model

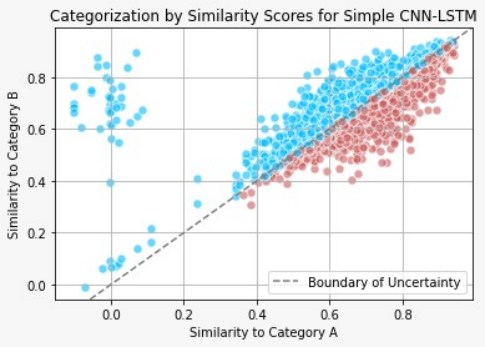


Figure 23: Similarity Scores for CNN-LSTM with Visual Attention model

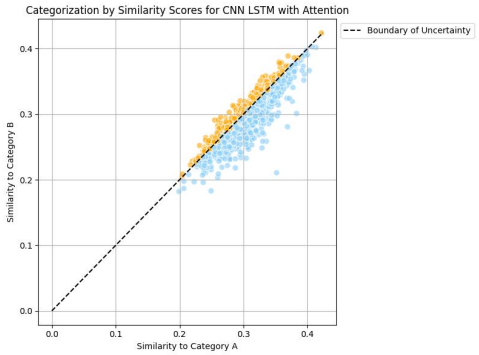


Figure 24: Similarity Scores for Vera

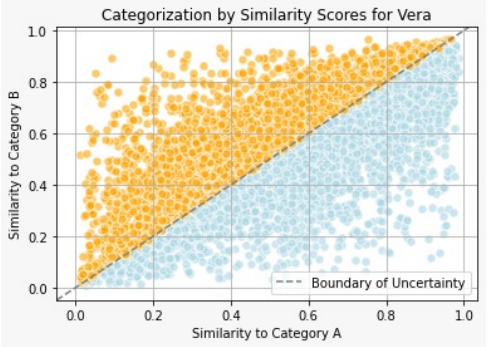
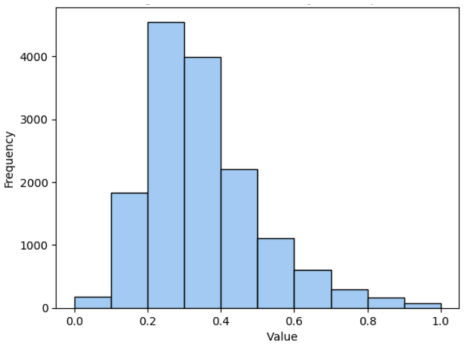


Figure 25: Probability Distribution of Selecting Category A (MLP)



Since the MLP model (see Figure 25) operates in a way entirely different to other models, by classifying each instance in the test dataset as either Category A or Category B. It does so by calculating sigmoid values. Due to this inherent difference, we see that the number of instances over which we calculate probability distribution doubles, however, the values of the mean and variance of this distribution still mean the same. For this particular baseline we see that our model is more varied in its confidence, and that it mostly predicts instances as not belonging to Category A.

For CLIP (see Figure 31) and CLIP Enhanced (see Figure 34), we see that the models mostly predicts that instances belong to Category A. However, we see that CLIP is more certain in its decision, whereas that is not the case with CLIP Enhanced.

For our final proposed model (see Figure 35), we see that the model is more confident about its predictions of whether an instance belongs to Category A, as compared to the unimodal and simple multimodal baselines. This trend is also seen in its comparison with competitive baselines like RN50, RoBERTa-ViT-B-32. From our proposed models, we only calculated this metric for our final proposal (Fusion and Shared Representation Learning Model ViT-B/32).

7.1.5 Image-Text Similarity

Since we are making use of a base image encoder as part of our pipeline, it is important to ensure that the embeddings generated by it, and the corresponding text embedding are as close as possible, before we perform any further operations on them. Since we are using only CLIP's text encoder, we are only measuring this metric for CLIP-based image encoders. Examples of the similarity scores of both encoders, for representative batches, can be seen in Figure 26 and Figure 27.

Figure 26: Image-Text Similarity scores for ViT-B/16

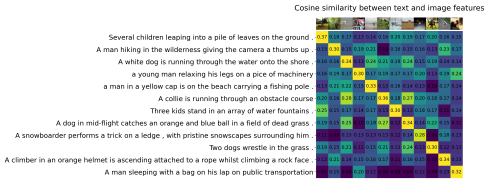


Figure 27: Image-Text Similarity scores for ViT-B/32

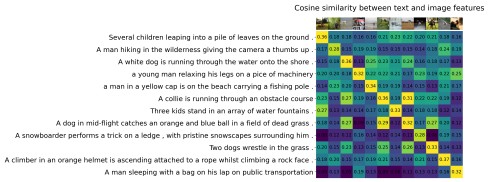


Figure 28: Probability Distribution of Selecting Category A (VisualBERT-based)

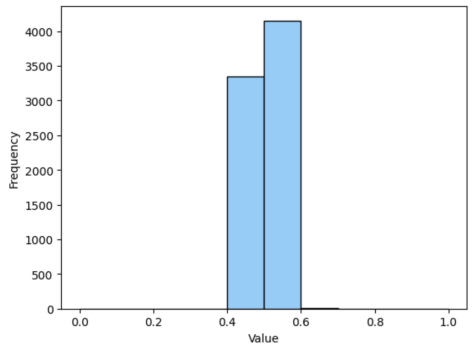


Figure 29: Probability Distribution of Selecting Category A (Vera)

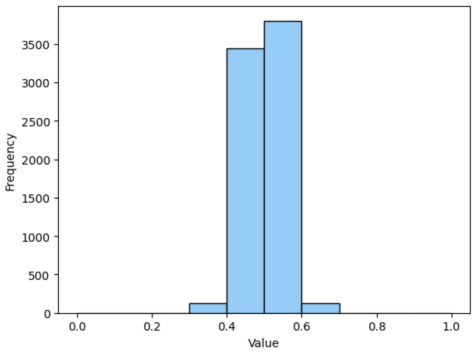


Figure 30: Probability Distribution of Selecting Category A (Simple CNN-LSTM)

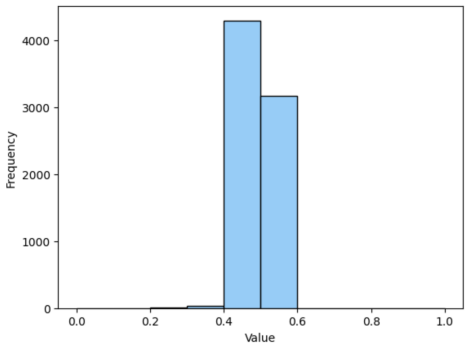


Figure 31: Probability Distribution of Selecting Category A (CLIP ViT-B/32)

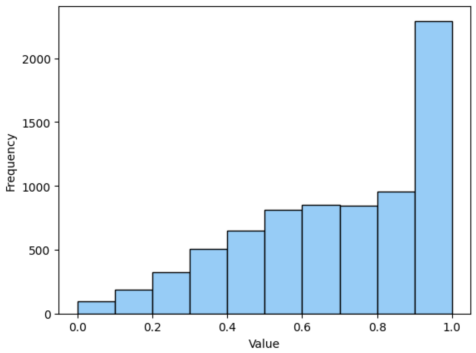


Figure 32: Probability Distribution of Selecting Category A (RN50)

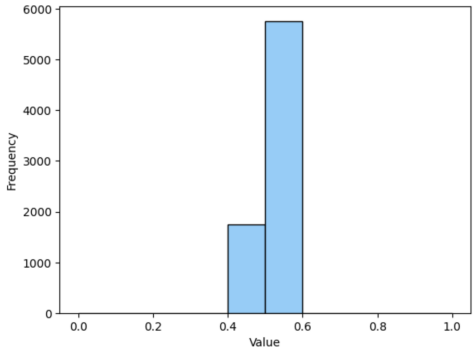


Figure 33: Probability Distribution of Selecting Category A (RoBERTa-ViT-B-32)

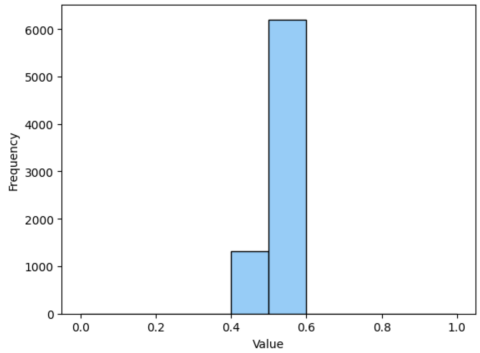


Figure 34: Probability Distribution of Selecting Category A (CLIP Enhanced-FineGrained)

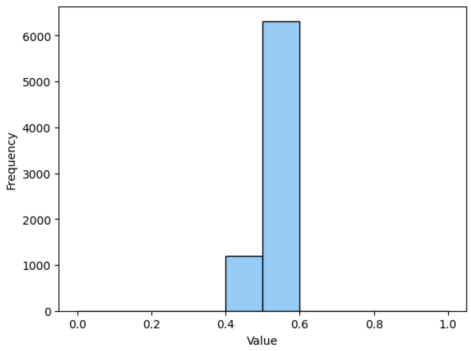
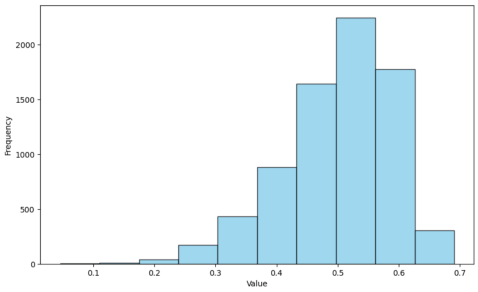


Figure 35: Probability Distribution of Selecting Category A (Fusion and Shared Representation Learning Model)



7.2 Qualitative Analysis and Examples

We now proceed to evaluate the some scenarios where the baseline models gave an incorrect output. The failures are with respect to the main task i.e. measuring the compositional understanding of vision-language models. We measure this by checking the accuracy with the model selects positive captions when compared against hard-negative captions (please refer to Report 2 for a detailed description of the task).

Failure cases for all the baseline cases can be seen in Figure 36. We have provided the corresponding image, the positive and hard-negative captions from which the model is expected to choose, and the scores assigned by the models to each of those captions. Since these are failure cases, the scores for the negative captions will be higher in all the instances. We should note that some models use probability values rather than similarity scores (like the MLP model, CLIP and CLIP Enhanced-FineGrained). We have chosen to group them all under "Scores" as well, for readability.

In all the cases, we can see that there are very subtle differences between the positive and hard-negatives, and therefore the models in these cases are unable to differentiate between them with any confidence. This can also be seen in the very small difference between the positive and negative scores in most cases.

An interesting aspect, is that these low difference can be seen in the more sophisticated RN50 and RoBERTa-ViT-B-32 (for both correct and incorrect classification).

The failures of the text-only models, MLP and Vera, are expected, since they are only looking at the plausibility of the statements, and the hard-negative captions are not blatantly unreasonable.

7.3 Insights

Our analysis of the baseline models' performance has helped clarify our path forward. It is evident that we need auxiliary loss functions for the image and text encoders in our proposed model (please refer to Report 2) that will minimize the distance between attributes and their corresponding entity nouns, as well as maintain a minimum semantic distance between positive and hard-negative caption pairs. This will help avoid scenario where our model might select the correct caption with a probability hovering around chance i.e. improve its confidence.

The importance of a good visual encoder has also been emphasized by our analysis. We would need to select encoders that provide embeddings that can differentiate between additional, swapped or replaced attributes and entities. We can see this difference between encoders by looking at the higher confidence with which CLIP ViT-B/32 makes its classification as opposed to an encoder like RN50. We need to make sure that we do not just focus on the overall accuracy of our model's performance, but also it's sub-tasks.

7.3.1 Failure case analysis for proposed models

The failure cases for our proposed models can be seen in Figure 37. The accuracy results can be found in Table 1.

Looking at the example, Model v1 clearly has difficulty in identifying swapped attributes/objects in the caption, indicating that the image and text modalities are not well aligned. This could be happening to the due to the attention module the text embeddings are being put through after we get them from the text encoder, which in our case is CLIP's. CLIP's text encoder will have been trained to provide embeddings aligned to the image embeddings received from its image encoder, and modifying the text embedding without applying cross-modal attention between the two after that would likely result in a poor similarity estimation.

Model v2 appears to be having difficulties in recongnizing nuances like quality. While both the statements "People sitting in the stands...", and "A man sitting in the stands....." are technically accurate descriptions, "people" is more apt in this scenario. There are other examples of such failures by this model. This would indicate that more work would be needed to make this model work with images containing crowded environments.

And our final model is unable to identify a object being added to the caption that is not present in the image. This example illustrates the point that even though adding an object would cause the shared embedding to vary significantly, which should allow the model to distinguish it more easily, the model is likely not deep enough to capture all the relationships between the image and the text. Additional cross attention modules should be explored.

8 Future work and Limitations

Our current model, despite good results, exhibits certain limitations in handling the nuanced inter-

play between textual descriptions and corresponding visual elements. The standard cross-modal attention mechanism that we are using, while effective for general tasks, struggles with the below issues:

- **Sensitivity to Contextual Ambiguities :** Even with advanced cross-modal attention mechanisms, the model may struggle to accurately interpret and align ambiguous textual descriptions with their corresponding visual elements. For instance, phrases with multiple valid interpretations can lead to mismatches between the predicted and actual visual contexts.
- **Lexical Sensitivity :** The model may lack the capability to fully understand and utilize the linguistic relationships within text, such as the connections between nouns and their modifiers, which are crucial for interpreting complex scenes. There is a scope of improvement in this area.
- **Robustness to Complex Compositional Structures :** While the model aims to enhance compositional reasoning, it may still be limited in handling highly complex compositional structures, such as nested or recursive visual relationships (e.g., "a painting within a painting" scenarios). The ability to parse and represent these multi-layered compositions could be constrained by the linear and attention-based mechanisms currently employed.
- **Handling of Polysemous Words :** The model may face challenges when dealing with words that have multiple meanings depending on the context (polysemous words). The ability of the text encoder to disambiguate such terms based solely on the immediate textual context might not always be effective, leading to incorrect image-text mappings.
- **Semantic Alignment :** The model at times may misalign semantic meanings between text and visual data, leading to inaccuracies when differentiating between visually similar objects that are contextually distinct.

To address these limitations, we propose replacing the standard cross-modal attention heads with LXMERT-based attention heads proposed by (Tan and Bansal, 2019). This modification aims to enhance the processing and integration of multimodal data, providing a more nuanced interaction between the text and image modalities.

LXMERT incorporates lexical bindings into the attention mechanism, making it inherently more sensitive to the specific linguistic structures and relationships within the text. This means that the model not only attends to areas of the image that are relevant to the text but does so in a way that respects the lexical relationships (like modifiers and nouns) defined in the text. This level of sensitivity is particularly beneficial when distinguishing between visually similar but contextually different images.

By using LXMERT, the model can better align semantic meanings between text and images. Lexicalized attention ensures that the model pays attention to the correct attributes of objects in images as described by the text, such as colors, sizes, or other specific properties that standard attention mechanisms might overlook or misalign.

Furthermore, in complex scenes where multiple interactions or attributes need to be understood and described, LXMERT can help by ensuring that the attention mechanism is not only focusing on the right parts of the image but also interpreting these parts in the context provided by the text. This is crucial for categorizing positive and negative image captions, especially when the image contains elements that are subtle or easily confused without textual guidance.

Another potential improvement for future work could be to incorporate the Cross-modal Hard Aligning Network (CHAN) described in (Zhengxin et al., 2023). This network specifically aims to enhance fine-grained image-text matching by eliminating redundant and irrelevant alignments.

CHAN leverages a novel approach to image-text matching that focuses on the most relevant region-word pairs and eliminates all other alignments. This method not only improves accuracy by ensuring that only the most semantically relevant pairs are considered but also enhances efficiency by reducing the computational load associated with cross-modal alignments.

CHAN can by integrated by replacing our current cross-modal attention mechanism with the

hard assignment coding approach described in CHAN. This would involve adjusting our attention mechanisms to focus exclusively on the highest weighted region-word pairs.

By focusing only on the most discriminative features between image and text, CHAN effectively reduces the influence of non-relevant alignments, which enhances the accuracy and efficiency of the retrieval process. Moreover, it is designed to ensure that the semantic correspondence between text and image fragments is more precise, directly addressing the issue of semantic alignment in our current model.

By adopting CHAN, we could enhance our model's capability to perform fine-grained image-text matching more efficiently and accurately, addressing both the precision and performance limitations currently faced.

9 Ethical Concerns and Considerations

In our pursuit to enhance the compositional reasoning capabilities of vision-language models (VLMs) for positive applications, it is imperative to acknowledge the potential dual-use nature of our technology. Just as advancements in AI can yield beneficial outcomes, they also carry ethical considerations that require careful attention. Here, we recognize several key ethical considerations:

- If the model learns and perpetuates biases present in the dataset, it could potentially reinforce or amplify harmful societal biases and stereotypes. For instance, if the dataset portrays certain gender roles or racial stereotypes, the model may learn and propagate those biases when applied to real-world scenarios, leading to unfair or discriminatory outcomes in areas like hiring, lending, or content moderation. Furthermore, our model aims to capture fine-grained details from the images and captions, and the impact of any bias in the dataset can be amplified by an effective model.

- The ability to accurately match images with captions and distinguish subtle differences could be exploited for unintended or malicious purposes. One specific concern is the potential for governments or authoritarian regimes to misuse this technology for censorship and content control. With the capability to search for and identify images based on their associated captions, governments could systematically locate and remove online content that they deem unfavorable or dissenting. This could infringe on freedom of expression and enable the suppression of diverse perspectives and criticism. For example, a government could use the model to search for and delete images or posts containing captions that criticize or disagree with their policies or ideologies. Even minor differences in the wording of captions could be used to identify and target content for removal. This kind of censorship would not only violate fundamental human rights but also stifle public discourse, limit access to information, and perpetuate government control over the dissemination of ideas. To mitigate this risk, it's crucial to implement robust access controls and auditing mechanisms to prevent the misuse of the technology by authoritarian regimes or bad actors. Additionally, clear ethical guidelines and governance frameworks should be established to ensure that the deployment of such models is transparent, accountable, and aligned with principles of free speech and freedom of expression. Measures such as encryption, decentralization, and community-driven oversight could also help protect against centralized control and censorship.

- Suppose the VLM is tasked with retrieving images based on the text prompt "a man in a prison cell." However, due to limitations in its compositional reasoning capabilities, the model fails to accurately distinguish the intended meaning and instead retrieves images depicting "a black man in a prison cell." This output not only fails to accurately match the original prompt but also perpetuates harmful racial stereotypes and biases. By associating the concept of a "man in a prison cell" with images of black individuals, the model inadvertently reinforces the deeply problematic and discriminatory narrative that disproportionately links black people with criminality and incarceration. Such biased associations can have severe consequences, particularly in decision-making contexts where the VLM's outputs may influence societal perceptions, policies, or practices. For instance, if this technology were integrated into crim-

inal justice systems or media platforms, it could reinforce existing racial biases, perpetuate discriminatory narratives, and contribute to the marginalization and dehumanization of certain communities. Moreover, the lack of awareness of social complexities and historical contexts within these models exacerbates the issue. VLMs may not fully comprehend the nuanced implications of their outputs, failing to recognize the potential harm caused by perpetuating harmful stereotypes or reinforcing societal biases.

Figure 36: Failure Cases for Baseline Models

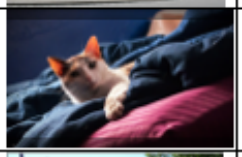
Model	Image	Positive Caption	Negative Caption	Positive Score	Negative Score
Multilayer Perceptron (Text only)		A man in black shirt standing in field with baseball mitt.	A man in baseball shirt standing in field with black mitt.	0.446	0.554
Simple CNN-LSTM		A table topped with a cake covered in berries next to a plate of sandwiches.	A crystal table topped with a cake covered in berries next to a plate of sandwiches.	0.618	0.622
CNN-LSTM with Visual Attention		The extremely small car is parked behind the bus	The bus is parked behind the extremely small car.	0.271	0.294
RN50		A kitten on a bad with its arm stretched out toward the camera	A kitten with a hat on its head is on a bed with its arm stretched out toward the camera.	0.223	0.241
RoBERTa-ViT-B-32		Two young women are washing two motorcycles with hoses.	Two young women are washing two gilded motorcycles with hoses.	0.327	0.330
Vera		a coin meter that has paint all over it	A coin meter that is clean all over it	0.369	0.558
CLIP ViT-B/32		A tall vase with red and white tulips in water.	A tall vase with red and white tulips next to a bowl of water.	0.492	0.508
VisualBERT-based model (Image only)		Blue bathroom with two white towels hanging by the shower.	White bathroom with two blue towels hanging by the shower.	0.190	0.194
CLIP Enhance-FineGrained		Child's bed and colorful quilt surrounded by blue plastic walls.	Child's bed and monochromatic quilt surrounded by blue plastic walls.	0.447	0.552

Figure 37: Failure Cases for Proposed Models

Model	Image	Positive Caption	Negative Caption	Positive	Negative
Basic Contrastive Learning Model (Proposed v1)		Two teddy bears a pink one, and a tan one. The tan bear is wearing a pink shirt that says the harvey girls.	Two teddy bears a tan one, and a pink one. The pink bear is wearing a tan shirt that says the harvey girls	0.260	0.294
Enhanced Cross Modal Attention Model (Proposed v2)		People sitting in the stands watching a man play tennis.	A man sitting in the stands watching people play tennis.	0.457	0.543
Fusion and Shared Representation Learning Model (Final)		A person that is cooking some food in a kitchen	A person and a parrot are cooking some food in a kitchen.	0.439	0.561

10 Team member contributions

Shrey Madeyanda contributed to various sections of the report, data pipeline setup, and to the implementation of the model and the loss function.

Shaurya Singh contributed to various sections of the report, and to the implementation of the image encoder, the model and the loss function.

Geerisha Jain contributed to various sections of the report, and to the implementation of the text encoder, hyperparameter tuning for text encoder, and qualitative analysis.

Madhura Deshpande contributed to various sections of the report, and to the data pipeline setup, qualitative analysis and experimenting with loss function.

References

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements.

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

Han Ma, Baoyu Fan, Benjamin K. Ng, and Chan-Tong Lam. 2024a. Vl-few: Vision language alignment for multimodal few-shot meta learning.

Teli Ma, Rong Li, and Junwei Liang. 2024b. An examination of the compositionality of large generative vision-language models.

Ziping Ma, Furong Xu, Jian Liu, Ming Yang, and Qingpei Guo. 2024c. Sycoca: Symmetrizing contrastive captioners with attentive masking for multimodal alignment.

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

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts.

Timothy Ossowski, Ming Jiang, and Junjie Hu. 2024. Prompting large vision-language models for compositional reasoning.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment.

Nermin Samet, Samet Hicsonmez, and Emre Akbas. 2020. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *European Conference on Computer Vision (ECCV)*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-modal attention with semantic consistence for image–text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5412–5425.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it?

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models.

Pan Zhengxin, Fangyu Wu, and Bailing Zhang. 2023. Fine-grained image-text matching by cross-modal hard aligning network. pages 19275–19284.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment.