Enhancing Multi-Modal Interactions in Memes for Downstream Tasks

Ajay Mittur* **Akshay Goindani*** **Ansh Khandelwal*** **Arushi Gupta***
{amittur, agoindan, anshk, arushigu}@andrew.cmu.edu

Abstract

In this project, we explore the Vision-Language (VL) interactions in memes. Previously proposed approaches perform well on VL tasks like Image Captioning, Visual Question Answering etc., where the information is either present in only a single modality, or both modalities contain redundant information. However, Memes are synergistic in nature i.e., the two modalities contain some information and help to interpret the meaning, but when integrated together, the meaning of the meme changes completely (e.g., contrasting information to induce humour). Through our evaluation of multiple approaches, we show that current state-of-the-art VL systems struggle to perform well on tasks related to Memes. Further, through our qualitative evaluation, we show that the tasks related to memes pose several challenges for these models, and in this work we focus on two challenges - need for external knowledge and complex reasoning. We use a Retrieval Augmented Generation (RAG) based approach to tackle the first problem, and fine-tune current state-of-the-art VL models on abductive reasoning datasets to tackle the latter challenge. Through our experiments on the MemeCap and Hateful Memes dataset, we show that ....

1 Introduction and Problem Definition

One of the core challenges towards creating robust Multimodal Machine Learning models is effectively capturing the interactions between the modalities (Liang et al., 2022). As per Liang et al. (2022), multimodal interactions have been categorized into three classes - 1) *Redundancy*, where the modalities have shared common information, 2) *Uniqueness*, where the information lies solely in one of the modalities, and 3) *Synergy*, where new information is emerged using the information from all the modalities (see Figure 1 for an example).

Figure 1: **Ground Truth Caption:** Meme poster appreciates their only two followers and one of them is their alternative account.

Previous works have shown that Vision-Language (VL) models have achieved great performance on tasks like Visual Question Answering (VQA) and Image Captioning, where *Redundancy* and *Uniqueness* are the prevalent interactions. However, the performance of VL models drop significantly when the interactions are synergistic (Yu et al., 2023). In this work, we focus on exploring ways to improve the performance of current VL models in the scenario of synergistic interactions. Particularly, we focus on understanding Memes as it not only requires object detection and language understanding, but also integrating the information from the two modalities to understand the visual metaphors. In order to study the multimodal interactions in memes, and to evaluate the performance of various VL models, we use the datasets - MemeCap (Hwang and Shwartz, 2023) and Hateful Memes (Kiela et al., 2020b). The latter dataset corresponds to the task of detecting whether a meme is hateful or not, whereas the former requires to generate a caption for the meme and is much more challenging.

We first evaluate various baselines that achieve great performance on VL tasks, including strong

*Everyone Contributed Equally – Alphabetical order

unimodal baselines such as LLaMA (Touvron et al., 2023). Through quantitative metrics such as BertScore, CLIPScore, Rouge and F1, we then identify models that are strong at fusing the information from the two modalities. We find that Q-former based fusion methods are much more stronger than the approaches that combine the modalities using linear/MLP layers. We next perform the intrinsic evaluation of the baseline models to strengthen our finding that Q-former based fusion is better than MLP based fusion. We also use the EMAP score (Hessel and Lee, 2020) as an intrinsic metric and find that various strong multimodal baselines do not learn complex multimodal interactions and are additive in nature i.e., the scores of the multimodal models are very similar to the model that just uses the additive operation on top of the unimodal outputs.

Next, through our qualitative evaluation of the best performing baselines, we find some major challenges posed by the tasks related to memes. In this work, we attempt to tackle two of the major challenges - need for external knowledge, and complex reasoning. We use Retrieval Augmented Generation (RAG) to provide external knowledge from a textual knowledge base. To achieve this, we first vectorize the knowledge base using the CLIP (Radford et al., 2021) text encoder. Next, for each sample in the evaluation set we use the meme's CLIP embedding to retrieve the relevant textual descriptions from the knowledge base. The retrieved descriptions are then passed to the pre-trained Vision Language Model (VLM) to generate the response. Previous works (Ramos et al., 2023a) show the benefits of using RAG for image captioning and it can be incorporated into any architecture without the need for training/fine-tuning. We also propose to improve the reasoning abilities of the model, by fine-tuning the pre-trained VLM on complex reasoning dataset such as SHERLOCK (Hessel et al., 2022). Pre-trained VLMs are often trained on image-text pairs where the text is a literal description of the image. However, in the scenario of memes, the image contains visual metaphors and the ground truth caption is not a literal image description. We experiment with the two proposed approaches separately and find that using RAG helps to improve the score on the caption generation task. However, we observed a drop in performance when evaluated on the Hateful Meme Classification dataset. For the fine-tuned models we observe improvements in some intrinsic metrics. However, overall the performance on the downstream tasks deteriorates.

In summary our contributions are as follows

1. Through quantitative and qualitative evaluation we show that Q-former based fusion methods are stronger than MLP based fusion.

2. In order to provide external knowledge to fixed models, we propose to use RAG for Meme Caption Generation and Hateful Meme Classification and show improvements on the Memecap dataset across all the metrics.

3. To improve the ability of the models to understand visual metaphors, we fine-tuned pre-trained VLMs for abductive reasoning using the SHERLOCK dataset.

2 Related Work and Background

Related Datasets The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes (Kiela et al., 2020b) introduces the Hateful Memes challenge, which aims to advance research on detecting hate speech in multimodal memes that contain both text and images. The authors release a new dataset called the Hateful Memes dataset containing over 10,000 meme images annotated as either "hateful" or "not hateful". Identifying hate speech in memes requires reasoning over the complex interplay between the visual and textual modalities. The authors describe several strong baseline models evaluated on this dataset for multimodal hate speech detection. A Visual-Only Model used a deep convolutional neural network trained solely on the meme images. A Text-Only Model employed a fine-tuned BERT model that took just the text transcripts as input. A Late Fusion Model combined predictions from the separate visual and text models using logistic regression. An Early Fusion Model concatenated visual and text features before the classification layer. Finally, a Visual-Guided Model incorporated full multimodal reasoning by allowing attention between visual and text representations in both directions. Out of these baselines, the Visual-Guided Model achieved the strongest performance by leveraging the combined semantics of the image and text. However, all models still underperformed human baseline accuracy, leaving ample room for developing more sophisticated multimodal techniques for the challenging hate speech detection task in internet memes.

LAION-5B (Schuhmann et al., 2022) is a massive open-source dataset containing over 5 billion image-text pairs, aimed at training next-generation image-text models. While its sheer scale is impressive, the dataset primarily consists of literal descriptions of images, lacking more complex reasoning or understanding tasks. Consequently, while models trained on LAION-5B may excel at image captioning and similar literal tasks, fine-tuning on datasets focused on reasoning, multi-step inference, and higher-level comprehension is likely necessary to unlock the full potential of these models for more advanced applications. Nevertheless, LAION-5B represents a significant resource for the vision-language community, providing a vast amount of data to pretrain powerful multimodal models that can then be adapted to specific downstream tasks through targeted fine-tuning. LAION-2B is a subset of this dataset.

Incorporating external knowledge for meme understanding is a crucial challenge, as memes often make references to popular culture, current events, and shared experiences that require contextual grounding beyond just the text and image. The paper by Hessel et al. (Hessel et al., 2023) introduces a novel dataset and suite of tasks derived from The New Yorker's weekly cartoon caption contest. A key contribution is the collection of dense annotations for each cartoon image, including descriptions of the scene, entities depicted, and explanations of what makes the scene unusual or humorous. These rich annotations serve as a proxy for the external knowledge required to fully comprehend the setup and punchline of each joke. Instead of having to learn and store all of this world knowledge from training data alone, RAG models could attend to the manually curated humor annotations at inference time when explaining memes.

The Abduction of Sherlock Holmes (Hessel et al., 2022) is a novel dataset designed to evaluate visual abductive reasoning capabilities of AI models. Abductive reasoning involves forming explanatory hypotheses based on observed evidence, a crucial skill for human-like intelligence. The dataset presents models with images containing anomalous or unexplained elements, and the task is to generate plausible hypotheses or explanations for these anomalies based on visual cues and commonsense reasoning. Unlike traditional vision-language tasks, this dataset goes beyond literal descriptions or question-answering by requiring models to make logical inferences and provide coherent explanations for visual oddities. This ability to reason about causality, formulate hypotheses, and bridge gaps in understanding is crucial for developing more human-like artificial intelligence systems. The Abduction of Sherlock Holmes dataset poses a significant challenge for current AI models, as it necessitates integrating visual perception, commonsense knowledge, and abductive reasoning skills – a combination that existing models struggle with. Consequently, this dataset serves as a valuable benchmark for evaluating the progress of AI systems towards more robust and generalizable visual reasoning capabilities.

Unimodal Baselines Table 1 shows the performance of unimodal baselines applied the task of hate detection on the hateful memes dataset (Kiela et al., 2020b) in existing work. TextBERT (Devlin J., 2019) applied only on the image captions gave the best AUC of 59% compared to the other unimodal models. On the vision modality, ResNet-152 (He et al., 2015) with a classification head gave the best accuracy of 64%. The hateful memes paper also established Image-Grid, which is an average pooled ResNet-152, and Image-Region – a Faster-RCNN architecture (Ren et al., 2016) with `fc6` layer finetuned using `fc7` weights.

Prior Work Supervised Multimodal Bitransformers for Classifying Images and Text (Kiela et al., 2020a) introduces supervised multimodal bitransformers (MMBT) - a novel architecture with two transformer encoders (one for image regions, one for text tokens) that attend to each other across modalities. Having dedicated encoders allows effectively capturing intra-modality relationships before cross-modal fusion. The bitransformer is trained on a combined loss for image classification, text classification, and an "alignmatch" loss that aligns representations of corresponding image-text pairs while pushing apart unrelated pairs. Experiments on multimedia benchmarks like MM-IMDB and Hateful Memes show bitransformers outperform previous unimodal and single-stream multimodal transformer models. They efficiently learn nuanced multimodal interactions critical for multimedia tasks. A key advantage is parameter sharing between the text and image encoders, enabling more scalable multimodal learning. The authors demonstrate bitransformers learn disentangled unimodal and multimodal representations.

Methods	Accuracy ↑	AUC ↑
TextBERT (Devlin J., 2019)	0.58	**0.59**
ResNet-152 (He et al., 2015)	**0.64**	0.50
Image-Grid (Kiela et al., 2020b)	0.50	0.52
Image-Region (Kiela et al., 2020b)	0.52	0.57

Table 1: Unimodal Baselines on Hateful Memes

Integrating Multimodal Information in Large Pretrained Transformers (Rahman et al., 2020) explores techniques for fusing multimodal information from images and text into large pretrained transformer language models like BERT and RoBERTa. The key challenge is effectively integrating the cross-modal signals during pretraining while retaining the rich unimodal knowledge acquired from self-supervised pretraining on text corpora. The authors propose two main methods: (1) Multimodal Mixture of Encoders (MMoE) and (2) Multimodal Adapters. MMoE involves an ensemble of transformer encoders - some operating only on text, others on both text and images. The outputs are combined with a gating network. Multimodal Adapters insert adapters (lightweight modules) at select layers of the transformer to integrate visual information. These approaches are evaluated by initializing models like BERT, RoBERTa, and ALBERT with pretrained multimodal weights and then fine-tuning on downstream multimedia tasks like visual question answering, multimodal reasoning, and multimodal sentiment analysis. Experiments show that both MMoE and Multimodal Adapters lead to significantly improved performance over baselines that are pre trained only on text. MMoE tends to work better for challenging multimodal tasks that require tight vision-language integration. Analysis attributes the gains to the multimodal models' improved grounding of visual concepts and ability to relate visual and linguistic representations. The authors also find the location of the adapters within the transformer is important.

The winner of the hateful memes challenge (Zhu, 2020) incorporated external knowledge into their classification model using Web Entity Detection and FairFace classifier. The proposed solution extracts relevant entities, races, and genders from the memes using these external sources which is then integrated into the visual-linguistic transformer framework, allowing the model to leverage these additional modalities. In particular, the author extends VL-BERT (Su et al., 2020) with entity, race, and gender tags to perform hate classification. Furthermore, the author explore techniques like repurposing the pre-trained image-text matching head for hate detection and ensemble modeling. The paper also suggests that incorporating knowledge graphs and enhancing pre-trained models with external knowledge could be a promising direction for improving hate speech detection in memes

Abductive reasoning is an essential quality for a model to be able to understand and reason over memes. Because cross modal interactions are not too obvious in memes, researchers have tried to distill knowledge from large language models (LLM) onto student models to help them perform better abductive reasoning. Lin et al. (Lin et al., 2023) propose an approach called MR.HARM to detect harmful memes by incorporating multimodal reasoning knowledge distilled from large language models (LLMs). They argue that existing harmful meme detection methods only capture superficial signals from the text and image modalities, failing to uncover the implicit meaning conveyed through their interplay. To address this, MR.HARM first conducts abductive reasoning with LLMs to generate rationales that explain the harmfulness of a meme based on the text, image caption, and label. These rationales are then used to fine-tune smaller language models in a two-stage training process, allowing the models to learn multimodal reasoning knowledge for better fusion of text and image representations. Through extensive experiments on three meme datasets, the authors demonstrate that MR.HARM achieves superior performance compared to state-of-the-art baselines by effectively leveraging the commonsense knowledge and reasoning capabilities of LLMs.

In an effort to improve the capability of models to perform abductive reasoning on memes, Sharma et al. (Sharma et al., 2023) introduce a novel task called EXCLAIM, which focuses on generating natural language explanations for the semantic roles (hero, villain, and victim) assigned to entities in memes. The authors propose a multimodal, multi-

task learning framework called LUMEN that leverages the complementary nature of semantic role prediction and explanation generation tasks. LUMEN incorporates visual, textual, and multimodal representations from various pre-trained models (ViT, DeBERTa, T5) to generate explanations for the assigned roles. The paper highlights the importance of multimodal cues and shared learning across related tasks for effectively generating explanations that capture the nuances and abstract reasoning present in memes. They emphasizes the need for abstract reasoning and contextualization to understand memes, which could potentially benefit from incorporating external knowledge sources.

Recent work however, has also shown tasks involving meme understanding to be more dominant on the textual modality and models performing better with natural language instructions. Suryavanshi et al. (Suryawanshi et al., 2023) propose an approach to transform the multimodal offensive meme classification task into a unimodal offensive text classification task. The authors argue that image features captured by encoders are unable to capture abstract representations like language, and therefore propose to leverage Natural Language Inference (NLI). They generate image captions using an off-the-shelf captioner and automatically transcribe the memes as if explaining them to a visually impaired individual. These transcriptions and labels are then transformed into an NLI format (premise-hypothesis-label). The authors demonstrate their approach by finetuning RoBERTa models previously trained on different tasks (emotion analysis, sentiment analysis, offensive tweet classification, and NLI) on three benchmark datasets (Memotion, Hateful Memes, and MultiOFF). Their approach achieves state-of-the-art results on the MultiOFF dataset and competitive performance on the other datasets, showcasing the potential of incorporating external knowledge through NLI for enhancing meme representations and improving hate speech detection in multimodal content.

In a similar direction, with better Vision-Language Models becoming generally available, researchers have been able to get good results using just natural language instructions. Cao et al. (Cao et al., 2023) explore using natural language prompts with frozen language models for the task of multimodal hateful meme classification. Rather than fine-tuning large pre-trained models, they investigate whether providing instructive prompts to models like CLIP and ALIGN can enable high performance on this challenging task in a more efficient and interpretable manner. They propose novel prompting techniques that incorporate information about both the text and image components of a meme. For the text, they use prompts that provide examples of hateful and non-hateful statements. For the images, they explore prompts that prompt the model to analyze certain visual attributes like symbols, objects and emotions present. By combining text and image prompts, their approach can leverage the pre-trained multimodal knowledge while injecting task-specific guidance. They evaluate on several hateful meme datasets and find their best prompting strategy achieves competitive performance to fine-tuned CLIP and ALIGN models while being significantly more efficient. Through analysis, they find their text prompts tend to focus on overt hate speech while the image prompts pick up on more subtle cues like offensive symbols. The text and image modalities provide complementary signals for meme hate classification.

Finally, Hee et al. (Hee et al., 2023) investigate the challenge of understanding the underlying hateful or offensive meaning conveyed in multimodal internet memes that combine images and text. The authors first curate a new dataset called HateMemes by filtering public meme repositories and annotating a subset as either hateful/offensive or not hateful, where the hateful memes target individuals based on protected characteristics like race, religion, gender, etc. They analyze the dataset to identify common coded hate speech techniques used in hateful memes, such as using visual metaphors, attacking minorities through majority in-group terms, and deploying multiple modalities to make the hate message more insidious. To benchmark computational models, the authors evaluate several multimodal classification approaches including a Visual-Only Model using image features, a Text-Only Model using text embeddings, a Concat-Based Fusion combining image and text representations, and an Attention-Based Fusion allowing cross-modality attention flows. The attention-based multimodal fusion model achieved the strongest performance by dynamically attending to the relevant hateful signals across both modalities. Additionally, the authors propose model interpretability methods to explain the hateful meme predictions, such as visualizing attention maps and retrieving neighboring training examples. The work sheds light on com-

mon coded hate techniques in memes and showcases the benefits of multimodal reasoning for detecting underlying hateful meanings, however, the results also expose current limitations in fully decoding the nuanced pragmatics of memes.

Relevant techniques InstructBLIP (Dai et al., 2023), the primary model behind our work, is an improved approach over BLIP (Li et al., 2022) that introduces "instruction tuning" to adapt a pre-trained vision-language model (BLIP) to follow instructions across various tasks in a zero-shot or few-shot manner. It employs a new attention mechanism called "Q-former" (Query Former) to better model interactions between the input image and instruction by generating task-specific queries used in cross-attention layers, allowing the model to attend to relevant image regions based on the given instruction. The authors evaluate InstructBLIP on diverse vision-language tasks like visual question answering, image captioning, and multimodal classification, demonstrating competitive or state-of-the-art performance and showcasing its general-purpose capabilities. Overall, InstructBLIP represents a significant step towards building general-purpose vision-language models that can adapt to a wide range of tasks with minimal fine-tuning.

UNITER is a state-of-the-art model proposed by (Chen et al., 2020) that learns universal representations for vision-and-language tasks by pre-training on large image-captioning datasets using novel pre-text tasks. It employs a Transformer architecture pre-trained with masked language/region modeling using conditional masking of one modality at a time. A key contribution is the Word-Region Alignment task that explicitly aligns words and image regions during pre-training via optimal transport. UNITER achieves new benchmarks across six V+L tasks like VQA, visual reasoning, and image-text retrieval by effectively fusing visual and textual semantics into joint embeddings. This capability of learning multimodal representations can be highly useful for meme analysis. Memes typically combine images with short texts conveying humor/commentary. By encoding meme images and texts into a shared embedding space, UNITER-like models can capture the semantic interplay between the visual and textual components, facilitating tasks like explaining meme intent/meaning or categorizing memes based on their multimodal content. The image-grounded text features from such models provide rich representations for building meme analysis systems.

Retrieval augmented pipelines are not new in this task. LMCAP (Ramos et al., 2023c) is a few-shot multilingual image captioning model that uses retrieval-augmented language model prompting without requiring any training on image captioning data. Given an input image, LMCAP first retrieves similar captions from a datastore using a multilingual CLIP model. These retrieved captions are then formatted into a prompt along with example captions demonstrating how to generate in the target language. This prompt is fed into a large multilingual autoregressive language model (XGLM) which generates the caption through few-shot prompting. LMCAP is inspired by retrieval-augmented generation approaches, but is the first to apply this in a multilingual captioning setting. Overall, LMCAP presents an efficient few-shot approach to multilingual captioning by prompting language models with retrieved knowledge, without expensive supervised multimodal training.

SMALLCAP, introduced by (Ramos et al., 2023b), is a lightweight image captioning model that leverages a retrieval-augmented generation (RAG) technique. It conditions caption generation on both the input image and a set of relevant captions retrieved from an external datastore, allowing it to exploit external knowledge and generalize better to new domains without expensive retraining or finetuning. The main components include a frozen CLIP vision encoder, a frozen GPT-2 language model as the decoder, newly introduced cross-attention layers connecting the encoder and decoder, an external datastore of text for caption retrieval based on similarity to the input image, and a prompt template that includes the retrieved captions as a task demonstration. The RAG technique could also benefit meme explanation by retrieving relevant text snippets from a broad knowledge base to provide the necessary context for understanding implied meanings, references, and jokes in memes, enabling more insightful and comprehensive explanations beyond just describing visual elements.

Captioning has also been used to enhance the joint representation of image and text in memes (Zhou et al., 2021). A sophisticated architecture combining image captioning and multimodal memes detection with key components being an Image Captioner, that extracting textual context from images, and an Object Detector, that identifies objects within memes. The Triplet-Relation

Network (TRN) forms the core, modeling cross-modality relationships among visual features, captions, and Optical Character Recognition (OCR) sentences. The Classifier then utilizes joint representations from the TRN to predict meme hatefulness. Through supervised and reinforcement learning, the architecture minimizes classification loss while maximizing the CIDEr score for image captioning, demonstrating superior performance in hateful memes detection. Extensive experiments conducted on multimodal meme datasets demonstrate the effectiveness of the proposed method, with promising results achieved in the Hateful Memes Detection Challenge.

Good knowledge bases are essential in RAG pipelines. To help models understand and classify hateful memes better, (Grasso et al., 2024) proposes a novel framework called KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection) that incorporates external knowledge into the process of identifying harmful memes. KERMIT constructs a knowledge-enriched information network for each meme by integrating entities extracted from the meme's text and images with relevant knowledge retrieved from the ConceptNet knowledge base. This meme knowledge graph captures relationships between entities as well as external commonsense knowledge that provides context for better understanding the meme's meaning. KERMIT then uses a dynamic learning mechanism with a memory-augmented neural network and attention to focus on the most informative portions of the meme knowledge graph for classifying whether the meme is hateful or not. Experiments on multiple hate meme datasets showed that explicitly modeling and integrating external knowledge into the classification process through KERMIT's knowledge-enriched information networks led to improved hate detection performance compared to various multimodal baselines that did not leverage external knowledge. The results demonstrated the importance of incorporating background knowledge and context for accurately interpreting and classifying harmful memes.

3 Task Setup and Data

In this work we use the MemeCap (Hwang and Shwartz, 2023) and Hateful Memes (Kiela et al., 2020b) datasets for our experiments and evaluations.

We first run strong unimodal baselines to show that both modalities are important for the task. Next, we implement/run some simple multimodal baselines by integrating the information from the two modailites trivially (e.g., addition/concatenation of feature vectors from pre-trained unimodal models). This helps us to evaluate various ways to integrate the information from the modalities and what kinds of pre-trained encoders are more useful for the task. Finally, we evaluate some competitive multimodal baselines on these datasets. This would further strengthen our observations and help us to make informed decisions. We quantitatively evaluate the models using the metrics (wherever applicable) - BertScore-F1 (Zhang et al., 2019), ROUGE (Lin, 2004), ClipScore (Hessel et al., 2021), Accuracy, and AUROC.

For the Memecap dataset, we use the train-val split (5823 samples) for training/fine-tuning any model. For evaluation, we use the test split (559 sammples). For Hateful Memes (Kiela et al., 2020b), we use a 75-25% train-val split. In our case, the test split is not really used since it's only purpose is for final evaluation in the datasets' competitions. The metrics used for both of these datasets are accuracy and area under receiving operating characteristics (AUC ROC) due the nature of the task, i.e. classification.

4 Baselines

4.1 Unimodal Baselines

In our datasets we found that understanding the memes require knowledge beyond what is already present in the image and text (e.g., past relations between TV show characters). Since Large Language Models (LLMs) have been pre-trained on a large amount of data (Touvron et al., 2023), they have access to vast explicit knowledge. Further, they can also generate fluent sentences. Hence, we evaluate LLaMA (Touvron et al., 2023) in the zero-shot setting, on the Memecap dataset. Moreover, we also provide LLaMA with some visual context by extracting the text written on the meme image using easyocr[1], and append the extracted text to the prompt. Hence, we evaluate LLaMA both with and without the OCR text. Following Hwang and Shwartz (2023), we use the following prompt to evaluate LLaMA.

We used the pre-trained LLaMA (7B) model from HuggingFace[2], and used the official weights.

[1] https://github.com/JaidedAI/EasyOCR
[2] https://shorturl.at/pGOV9

```
<image>This is a meme with the title "{title}".
The image description is "{image caption}".
The following text is written inside the meme:
"{OCR text}".
```

Figure 2: Prompt for LLaMA evaluation

This is one of our strong unimodal language baselines. We evaluate LLaMA only on the Memecap dataset, because for the other two classification tasks, the generations were free form and verbose which made the evaluation based on Accuracy and AUROC challenging.

For the classification task on the hateful memes dataset and memotion dataset, we used a BERT-based model (Devlin J., 2019). BERT is a robust choice as an unimodal baseline for classifying hateful memes owing to several factors. Firstly, it excels in semantic comprehension, capturing nuanced meanings and context inherent in textual content, which is crucial for identifying hate speech. Additionally, its pre-trained representations offer a strong foundation, allowing for efficient fine-tuning on specific tasks, even with limited labeled data. Moreover, its proven state-of-the-art performance across various NLP tasks underscores its effectiveness. After incorporating the necessary adjustments to the BERT model for the sentence classification task, the [CLS] symbol was added to the input sequence. This modification enables BERT to process statement-level input effectively, with [CLS] serving as a representation of the text's semantics.

Our unimodal visual baseline for classification was ResNet. ResNet is a highly effective model for image classification due to its innovative use of residual learning, that facilitate the training of very deep neural networks by mitigating the vanishing gradient problem. Moreover, it enhances parameter efficiency, resulting in impressive performance with fewer parameters compared to alternative architectures. Moreover, it is easily adaptable, hence, a reliable and versatile choice for a wide range of unimodal image classification tasks. Thus, allowing us to achieve good results on different types of classification tasks in our datasets.

4.2 Simple Multimodal Baselines

For simple multimodal baselines, we used pre-trained vision and language models and fused them using a linear-layer/MLP to generate multimodal representations and use those for the task. We choose this setup as it will allow us to evaluate whether a simple fusion of modality representations can help understand synergic interactions. For the caption generation task, we use a transformer (Vaswani et al., 2017) based vision encoder and a causal Language Model (LM) to generate captions. We choose a Vision Transformer (ViT) based encoder as opposed to a ResNet model, as ViT based models are better at image captioning as compared to CNN based architecutres (Elbedwehy et al., 2022). Using the pre-trained vision encoder, we first get the image features. Next, we apply a learnable linear layer to transform the image features to the LM dimensions. Finally, the causal LM uses the transformed feature and attention mechanism to generate the caption. During training, the pre-trained image encoder and LM are frozen and we only train the linear layer used to transform image features. We train the model by optimizing the cross-entropy loss using parallel image-caption pairs.

In our experiments we used the pre-trained Vision Transformer (ViT) model `google/vit-base-patch16-224` from HuggingFace as the vision encoder and used `openai-community/gpt2` as the Causal LM. We call this model as ViT + GPT2. Further, since vision encoders pre-trained with linguistic supervision perform better on downstream tasks (Merullo et al., 2022), we also experiment with pre-trained Clip vision encoder as our image encoder. We call this baseline as Clip + GPT2. The choice of ViT/Clip in our experiments, is attributed to their large amount of pre-training data and robust downstream performance on various tasks. The Clip encoder used in our experiments is `openai/clip-vit-base-patch32` from huggingface.

For classification tasks, we used a Visual-BERT-based classifier (Liunian Harold Li, 2019). Visual-BERT is a novel multimodal framework that integrates the BERT language model with Faster R-CNN, an object detection model, to address tasks requiring analysis of both text and image inputs. Similar to its predecessor, the textual input structure of Visual-BERT-based classifiers mirrors that of BERT-based classifiers. However, it extends this structure by introducing additional embedding layers. While BERT utilizes three embedding layers for each subword, Visual-BERT enhances this by incorporating a visual feature embedding layer.

Furthermore, Visual-BERT introduces a special token, "[IMG]", along with its corresponding token embedding, obtained through the fusion of Faster R-CNN model outputs.

For text tokenization, the `bert-base-uncased` tokenizer is employed. Additionally, the framework utilizes the `vit-base-patch16-224-in21k` pre-trained model to extract vision features. The pre-trained Visual-BERT model `uclanlp/visualbert-nlvr2-coco-pre` serves as the foundational model for the fine-tuning process.

4.3 Competitive Baselines

Recently, pre-trained Large Vision-Language Models (LVLMs) have achieved state-of-the-art performance on various multimodal tasks. Further, they also allow the functionality of in-context learning, chain-of-thought prompting, etc. With such robust pre-training on large amounts of data and tasks, LVLMs become a strong and competitive baseline for our work. Hwang and Shwartz (2023) evaluated MiniGPT4 (Zhu et al., 2023) on the Memecap dataset. However, in this work we evaluate InstructBLIP (Dai et al., 2023) which is better at generating captions as compared to MiniGPT4, has a more complex fusing module - a Q former (Li et al., 2023), and does not rely only on the cross-entropy loss for training. We evaluate the performance of InstructBLIP in the zero-shot setting, and similar to LLaMA we uses the prompt (Figure 2), with and without the OCR text.

For the sentiment classification task in Hateful Memes and Memotion, we ran two open source multimodal models. The first was MMBT (Kiela et al., 2019) developed by Facebook and the second was Memotion Multimodal Model (M4) (Terence, 2021). Both the models used pre trained weights which were fine tuned on the datasets. Memotion consists of a multiclass classification task for the sentiment whereas both of these models formulate the problem as a binary classification task, hence to accommodate the models, the five sentiments in memotion (very positive, positive, neutral, negative, very negative) were binned to two categories, positive and negative (including neutral). The metrics were then calculated over these binary labels.

MMBT is a model whose key insight is that a transformer with unary self-attention over both text and image features, with a simple projected embedding alignment, can achieve highly effective multimodal fusion and classification without needing complex pretraining objectives. It combines pretrained BERT text encoder and ResNet image encoder, and learns linear mappings to project image features into the BERT token embedding space. This allows the model to attend jointly to both text and images. The model is also robust to missing images at training time, much more so than a late fusion baseline.

M4 is a model whose key insight is that transfer learning can be leveraged on a language model trained on a large corpus of text that usually contains high levels of sarcasm, humour, offense, and motivation to improve the encodings of meme captions and in turn improve the multimodal image and caption representation. It combines image embeddings obtained from VGG16 and text embeddings from ALBERT trained on Reddit data to get a multimodal representation of the meme. This vector is then passed through a Gated Multimodal Layer which decides the multimodal features to pass through to the prediction network. The prediction network consists of a simple 2 unit classification layer pre-normalized with batch norm and dropout.

5 Proposed Model

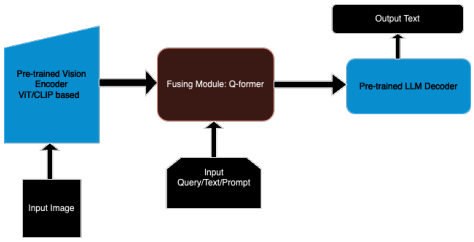

Figure 3: Initial Model Structure

Our quantitative results show that Large VLMs like InstructBLIP that use Q-former based fusion outperform other models that use MLP based fusion. Hence, in this work we propose to use InstructBLIP as the backbone for our model. InstructBLIP is also beneficial as the Q-former is pre-trained using three loss functions - Image-Text Matching loss (ITM), Image-Text contrastive loss (ITC), and the auto-regressive text generation loss. The ITM and ITC loss functions help the Q-former learn complex fusion functions. Based on the quantitative results, the proposal of the model is shown in Figure 3.

From our qualitative evaluation, we found that external knowledge is needed to perform well on tasks related to memes, as fixed models might not have all the information needed to understand the relationship between the objects in the meme and their metaphorical role. Previous works have used RAG to provide external knowledge to LLMs (Gao et al., 2023) and there has also been work on using RAG for image captioning (Ramos et al., 2023a). In addition, RAG can be applied to small models to achieve strong performance compared to using Large Models (Ramos et al., 2023a). In our qualitative evaluation, we also found that InstructBLIP focuses more on visual signals and generates the OCR text from the memes frequently. Hence, we hypothesize that by providing more textual context to InstructBLIP, it will be able to focus on the textual information as well.

InstructBLIP + RAG: In this project, we construct a textual knowledge base for RAG. We first curate relevant textual descriptions (more about the sources in the next subsection). Next, we vectorize the textual descriptions using CLIP's text encoder (Radford et al., 2021) and store the vectorized knowledge base for RAG. At the time of inference, the input meme is encoded using CLIP's image encoder. We use the cosine similarity between the encoded image and vectorized textual descriptions to retrieve the top-K similar descriptions. Finally, the retrieved K descriptions are used in the prompt for InstructBLIP for generating the relevant response. We use CLIP embeddings for retrieval. However, other multimodal embeddings can also be used and we leave the exploration of better embedding models for future work. Our proposed architecture is shown in Figure 4. Note: our approach does not require any re-training/fine-tuning and can be incorporated with any generative model.

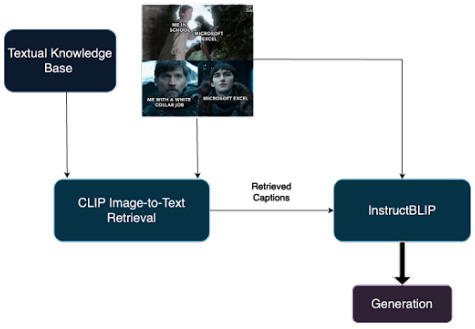

Figure 4: Proposed Model1: InstructBLIP with RAG using CLIP based embeddings

InstructBLIPFineTuned: From our literature review, we also found that the Image-Text pairs used to train current VL models do not provide sufficient incentive to the models to learn complex reasoning skills (Hessel et al., 2022). We hypothesize that by pre-training/fine-tuning VL models on complex reasoning tasks like SHERLOCK (Hessel et al., 2022) can help the models learn better multimodal representations for synergistic interactions, as these tasks require to utilize information from both modalities and reason about the emerging information as well. In this project, we choose the SHERLOCK dataset as it involves samples where abductive reasoning is required to perform well. The task for the models is to reason about the given image and a clue and generate an appropriate inference of the given situation. In cases of memes, such reasoning is essential to understand the metaphorical roles of different visual objects. Hence, we fine-tune the InstructBLIP model on the SHERLOCK dataset. To make the fine-tuning task more difficult, we remove the input clues provided in the dataset, and the model only looks at the input images to generate the inference.

Note: for the scope of the project, we evaluate the two proposed approaches individually. However, since our proposed RAG method can be used with any model, the two approaches can be applied together in future.

5.1 Loss Functions

Our first proposed approach i.e., InstructBLIP + RAG, does not require any fine-tuning/pre-training. Hence, we do not have any new loss functions in addition to the ITM, ITC, and auto-regressive loss, which is used during the pre-training of InstructBLIP. For our second proposed method, we use auto-regressive generative loss to fine-tune the InstructBLIP model. Since the Q-former is responsible for learning the fusion between the modalities, we perform the full fine-tuning of the Q-former and freeze all other parameters of the model. For future work, we would also like to explore LoRA (Hu et al., 2021) based fine-tuning of the Q-former architecture.

5.2 Knowledge Base and Sources for RAG

Memes are created from very diverse sources and curating data from all these sources is infeasible due to various reasons. The major reason being the accessibility of such datasets. For example, a lot of memes refer to some movie/TV-show charac-

ters, having access to the data describing the details about all characters/scenes is out of the scope of this project. Hence, we construct our knowledge base using open-source datasets that are most relevant to our downstream tasks.

For the Hateful Memes dataset, we leveraged the manually annotated dataset introduced in Decoding the Underlying Meaning of Multimodal Hateful Memes (Hee et al., 2023). The authors of that work curated this dataset by performing manual annotations that explicated the underlying contexts behind each meme instance deemed hateful within the Facebook Hateful Memes dataset. Given our utilization of the same dataset, we exercised prudence to avoid incorporating contextual annotations corresponding to the specific split on which our model was undergoing evaluation. This approach aimed to maintain the integrity of the evaluation process and avoid any potential data leakage that could artificially inflate model performance.

We also curated textual captions from the TV Captions dataset (Lei et al., 2020). We use this in the knowledge base, as from our evaluation we encountered a lot of memes referring to characters and scenes from TV shows. Ramos et al. (2023c) uses the COCO captions dataset as their knowledge base. However, since COCO captions dataset contains generic captions, COCO captions based knowledge base is not useful to our task.

5.3 Changes to training data

For the InstructBLIP + RAG approach, we do not have any training procedure. Hence, we do not make any changes to the training data. For the second proposed approach, where we fine-tune the model on the SHERLOCK dataset, we choose to perform instruction tuning similar to Dai et al. (2023), and modify the inputs accordingly. Specifically, we use the prompt *Q: What can you infer from this image? A:* and the image from the SHERLOCK dataset as input to the InstructBLIP model and optimize the autoregressive cross-entropy loss with the ground truth. As mentioned above, we do not use the clues provided in the dataset to make the task harder. However, the clue can be easily incorporated in the instruction. Since the InstructBLIP model is large and due to limited compute, we use the validation data of the SHERLOCK dataset for fine-tuning. We randomly split the validation set with 10% of the validation set being used as the test set.

5.4 Hyperparameters and their effects

For the InstructBLIP + RAG approach, the number of retrieved captions - K, is a very important hyperparameter. A lower value of K will not provide enough information to the model to generate the appropriate response. On the other hand, a higher value of K might include irrelevant context into the the prompt which can lead to inappropriate generations. We experiment with various values of K and show the results.

The temperature used for generation in generative models is also a very important hyperparameter as lower values of temperature can restrict the model from generating more meaningful responses. However, higher values of temperature can make the model generate incoherent responses. Through the experiments on the baseline InstructBLIP models we found that the temperatrue value of 0.3 works best consistently, and use that for our proposed models.

For our second proposed approach, we fine-tune the InstructBLIP model. Since these models have a lot of important hyperparameters such as Learning Rate, Batch Size, Number of warm up steps, varying them is infeasible and for fine-tuning we refer to the standard hyperparameter as per (Dai et al., 2023). We fine-tuned the InstructBLIP model for 4 epochs with a learning rate 1e-5, warm up steps of 500, per-device batch size of 64, and a weight decay of 0.01. We use a single A100 GPU for fine-tuning.

6 Results

The results for the caption generation task are presented in the Table 2. Note that ViT + GPT2 and Clip + GPT2 are trained using the Memecap training data, and other models are pre-trained and evaluated in the zero-shot setting. We used the following metrics to evaluate the performance.

ROUGE ROUGE evaluates the similarity between the generated text and the ground truth using n-gram overlap. ROUGE-L compares the Longest Common Subsequence between the two sentences. A higher ROUGE score is better as it tells that the generated text is structurally similar to the ground truth. However, it does not necessarily mean that the generated sentence is semantically similar to the ground truth. Based on the results, InstructBLIP + RAG (k=5) achieves the highest ROUGE score. ViT + GPT2 also achieves a high ROUGE

Model	BertScore-F1	ROUGE-1	ROUGE-2	ROUGE-L	CLIP Score
LLaMA with OCR	0.86	0.22	0.06	0.19	0.71
LLaMA without OCR	0.84	0.17	0.03	0.14	0.68
Clip + GPT2 (fine-tuned)	0.85	0.19	0.06	0.18	0.66
ViT + GPT2 (fine-tuned)	0.86	0.25	0.13	0.24	0.64
InstructBLIP without OCR	0.87	0.18	0.05	0.15	0.77
InstructBLIP with OCR	0.87	0.20	0.07	0.17	0.77
InstructBLIP + RAG (k=5)	**0.88**	**0.30**	**0.20**	**0.32**	**0.79**
InstructBLIP + RAG (k=1)	0.87	0.20	0.07	0.10	0.77
InstructBLIP + RAG (k=3)	0.87	0.28	0.15	0.26	0.77
InstructBLIPFineTuned	0.84	0.15	0.06	0.14	0.72

Table 2: Performance of Various Baselines and proposed models on the MemeCap test set

	Dev	
Methods	Accuracy ↑	AUC ↑
TextBERT (Devlin J., 2019)	0.58	0.59
ResNet-152 (He et al., 2015)	0.64	0.50
VisualBERT (Li et al., 2019)	0.60	0.58
MMBT (Kiela et al., 2020b)	0.63	**0.68**
M4 (Terence, 2021)	0.55	0.55
InstructBLIP	**0.67**	-
InstructBLIP + RAG (k=5)	0.63	-
InstructBLIP + RAG (k=3)	0.59	-

Table 3: Comparison of RAG InstructBLIP on Hateful Memes

score, however, qualitatively, the generations of ViT + GPT2 are not meaningful with respect to the meme. Further, since the gold captions in the Memecap dataset has the most frequent substring - "Meme Poster is trying to convey", by training the model on the training set, the model achieves a high ROUGE score by just generating the substring "Meme Poster". Therefore, ROUGE is not sufficient and reliable for complete evaluation, and hence we use BERTScore as well.

BertScore BertScore evaluates the semantic similarity between the generated text and the ground truth. This helps us to evaluate whether the generated text is meaningful and relevant. The highest BERTScore is achieved by the InstructBLIP + RAG models. There is no difference in the performance of InstructBLIP with and without OCR, this might be due to the ability of InstructBLIP to encode the images well and hence, the OCR information in the prompt is redundant. Also InstructBLIP + RAG (k=1) achieve similar scores as compared to InstructBLIP without RAG. This could be due to the insufficient information provided by just a single caption. Further, LLaMA also achieves a BertScore of 0.84, and when provided with the OCR information, the score is similar to that of InstructBLIP. All the metrics improve when OCR details are provided to the models, this implies that visual information is necessary to perform well in this task. From the generations of the models, we observed that the high BertScore-F1 is inappropriate, and the score computed is highly dependent on the pre-trained LM used to evaluate. The relative difference between the models is also contingent. In these results, we use the default model: `roberta-large` , from huggingface. In the task of image-captioning, the generated caption should also be relevant to the input image, ROUGE and BertScore-F1 only measure the relevance with the ground truth. In order to evaluate the relevance with the input image, we next use the Ref-CLIPScore (Hessel et al., 2021).

CLIPScore Ref-ClipScore uses the CLIP embeddings of the image, predicted text and the ground truth using the image and text encoders and calculates the relative similarity score similar to F1-

score. InstructBLIP + RAG achieves the highest CLIP score followed by LLaMA with OCR. Further, CLIP + GPT2 performs better than ViT + GPT2, and the generations are more relevant as compared to the latter.

Accuracy The accuracy of the baselines we ran are shown in Table 3. It gives us a measure of how many memes were correctly classified out of all the memes in the dataset. The unimodal visual model, ResNet achieved an accuracy of 0.64 on Hateful memes and 0.52 on Memotion. The unimodal text-based BERT model attained an accuracy score of 0.58 when evaluated on the Hateful Memes dataset. In comparison to this, the performance of Visual BERT did not exhibit a significant improvement. However, upon evaluation of the Memotion dataset, a discernible elevation in accuracy score was observed with Visual BERT registering at 0.59, contrasting with the performance of Text BERT, which yielded a score of 0.46. The MMBT model achieved an accuracy of 0.63 on Hateful memes and 0.52 on Memotion. The M4 model achieves similar accuracies as MMBT, on the Memotion datasets. The difference in accuracy is almost negligible between both the models which indicates pretraining on a large corpus of data from social media may not be as helpful as expected. The M4 model uses transfer learning from an ALBERT model trained on Reddit data to encode the meme captions. The approach used in MMBT – projecting the image features onto the token embeddings, seems to work as well as using a model pretrained with social media knowledge. But altogether, linear mappings of textual encodings over image features does not result in overly impressive accuracies either.

InstructBLIP with a Q-former based fusion module achieves the highest accuracy. However, with RAG the performance of the model deteriorates. Upon further evaluation of the retrieved captions, we found that the retrieval knowledge based for memes related to hateful categories is not very informative. Hence, the quality of the knowledge base is a major limitation towards the performance of our proposed approach.

Area Under Curve - Receiver Operating Characteristics (AUROC) The AUROC scores for the baselines that we ran are also shown in Table 3. AUROC provides a comprehensive assessment of a binary classification model's ability to make correct predictions across different decision thresholds. It is a widely used metric because it is threshold-independent and provides insights into the overall discriminatory power of the model. Kiela et al. (2020b) also recommends to report the area under the receiver operating characteristic curve as the main metric. All the models achieve similar AUROC scores on the memotion dataset which implies that the models have a low discriminatory power for this task. For Hateful memes, the AUROC scores were higher. The MMBT model achieves the highest AUROC score of 0.68, this can be attributed to the large-scale pre-training of the unimodal components. Note: we do not calculate the AUROC scores for the InstructBLIP based model as the free form generation poses challenges towards interpreting the probability/confidence scores.

7 Analysis

7.1 Intrinsic Metrics

In this section we evaluate the models on the basis of some skills essential to solve the tasks related to Memes. One of the most important skill required to understand memes, is the ability to effectively combine the information from the two modalities. In order to evaluate whether the models are effectively utilizing the information from the modalities, we evaluate the following intrinsic metrics. The results of our evaluations are presented in Table 4. For one of our baselines, ViT + GPT2, we were unable to evaluate any intrinsic metric, as the metric computation is not compatible with the architecture of the model. Also, for our proposed InstructBLIP + RAG, since we do not perform any fine-tuning, the intrinsic metrics are the same as the base InstructBLIP model.

Image-Text Matching (ITM) helps to measure the fine-grained alignment between image and text. Fine-grained alignment is essential for utilising the information from the modalities as entities in one modality often refer to the entities in the other. Hence, this metric is very useful for our task. In our work, we use this metric only for the InstructBLIP (Dai et al., 2023) model as it has a separate binary classification head trained to determine whether the image and text match or not. Since the match between an image-text pair is not formally defined, we do not use this metric to evaluate other models.

The binary classification head for ITM calculates the probability of a match between the corre-

Model	ITM	ITC	Hit@10	Hit@5
CLIP + GPT2 without Linear Projections	-	0.29	77.3	72.1
CLIP + GPT2 with Linear Projection	-	0.78	35.4	28.4
InstructBLIP	0.79	0.65	87.4	83.1
InstructBLIPFineTuned				
VisualBERT without Projections	-	0.24	-	-
VisualBERT with Projections	-	0.42	-	-

Table 4: Performance of different models on our Instrinsic Metrics. InstructBLIP based models and CLIP + GPT2 are evaluated on the Memecap test set. VisualBERT is evaluated on the hateful memes dataset for the classification task.

sponding image-text pairs. For the InstructBLIP model, using the image-caption pairs in the Memecap test set, the mean ITM score is 0.79. Since the mean ITM score is significantly higher than 0.5, this implies that InstructBLIP is good at capturing the fine-grained alignment between Image-Caption pairs. For the InstructBLIPFineTuned model, the ITM score is 0.65, which is less than the ITM score of InstructBLIP. This might be one of the reasons behind the lower performance of InstructBLIPFineTuned.

Image-Text Contrastive Learning (ITC): In late fusion based models, the representations obtained from different modalities are first projected to a common multimodal space and then the multimodal representations are used in the downstream tasks. By training the model to optimize the objective on the downstream tasks, the projection layers are learned to appropriately project the unimodal features into the multimodal space. Image-text contrastive learning allows to measure how well the image and text representations in the multimodal space are aligned. In order to measure this, similar to Li et al. (2023) we compute the cosine similarity of the ground truth image-caption pairs using their multimodal representations. A higher similarity score implies better aligned representations. We evaluate this metric only for the CLIP + GPT2, InstructBLIP, and VisualBERT models, as only these models among our baselines have a late fusion component. InstructBLIP uses a Q-former (Li et al., 2023) to create a multimodal space, whereas in the CLIP + GPT2 model, we use linear projection layers to project the image and text representation from the CLIP encoders to create a common multimodal space. VisualBERT employs a transformer-based architecture specifically designed to handle multimodal inputs to create a latent projection space for the text and visual embeddings.

The ITC for the InstructBLIP model on the Memecap Test set is 0.65 and that of the InstructBLIPFineTuned model is 0.67. The increase in the ITC score indicates that the image and text representations have become closer. For the CLIP + GPT2 model, we first calculate the ITC using only the CLIP encoders, which comes out to be 0.29. Next, we apply the learned projection layers to these encodings to create the multimodal representations and again compute the ITC of the multimodal representations. The ITC using the projected representations is 0.78. This increase in the ITC score implies that the simple linear projection layers are effective at improving the similarity between the modalities and help in learning good multimodal representations.

We evaluated the VisualBERT model on the hateful meme dataset for classification. Initially, the ITC score using separate BERT text encoder and ViT image encoder was 0.24. After projecting these representations into a shared embedding space, the ITC score increased to 0.42. While there was an improvement, it was less substantial compared to the CLIP + GPT2 model, indicating that CLIP + GPT2 aligns embeddings more effectively. The ITC achieved by the CLIP + GPT2 model is better than the InstructBLIP model. However, ITC only captures the similarity between the matching pairs. We next design an intrinsic metric that also helps us to evaluate the similarity between the unmatched image-caption pairs.

Hit@K: ITC measures the closeness of the ground-truth image-caption pairs in the multimodal space. However, this is insufficient as it doesn't measure the similarity between the image-caption pairs where the caption is not relevant to the im-

age. For a good multimodal space, the similarity between such pairs should be low and the similarity of the ground truth pairs should be high. Further, such multimodal spaces have been widely used for image-text retrieval (Radford et al., 2021). With good multimodal representations for both image and text, the performance of cross-modal retrieval is improved. Hence, we use cross-modal retrieval performance of the learned embedding space to evaluate whether the unmatched image-caption pairs are distant or not.

To achieve this, we use the Hit@K metric which calculates the percentage of images for which the relevant ground truth caption lies in the top K most similar (cosine similarity) captions retrieved. A low value of Hit@K implies that for a lot of images, the corresponding ground truth caption is not among the top K retrieved samples and the multimodal space learned isn't effective. For evaluating Hit@K, we use the memecap test set which provides us with image-caption pairs, and we only evaluate the CLIP + GPT2 and InstructBLIP models for this metric.

Similar to ITC, we compute Hit@10 for the CLIP + GPT2 model in two stages - with and without the linear projection layers. The Hit@10 for CLIP without linear projection is 0.77 whereas the Hit@10 with linear projections is 0.35. The Hit@10 achieved by InstructBLIP and InstructBLIPFineTuned are 0.87 and 0.81. The drop in the Hit@10 indicates that although the fine-tuning helped to bring the modalities closer in the space, the unmatched pairs are also closer which deteriorates the retrieval performance. Similarly, although, ITC for CLIP + GPT2 improved with the use of projection layers, a significant drop in Hit@10 suggests that the similarity between irrelevant image-caption pairs also improved with the linear layers which affects the retrieval performance. A high Hit@10 score for InstructBLIP provides an evidence that the multimodal representation learned using the Q-former is meaningful and better aligns the image-text representations as compared to the linear layers used in CLIP + GPT2.

EMAP: Empirical Multimodally Additive Projection (EMAP) (Hessel and Lee, 2020) is diagnostic tool that allows to evaluate whether the models (that use cross-modal interactions, also referred to as interatctive models) actually learn cross-modal interactions. A pre-trained multimodal model is projected to a less-expressive space which basically resembles the additive multimodal space. Additive multimodal models are ensembles of respective unimodal models combined using the addition operation. Such models learn little to no interactions between the modalities. Using EMAP we first project the pre-trained multimodal models to an additive space. Next, we evaluate the metrics such as Accuracy, F1-score, etc., for the projected models. If the performance of the EMAP projected model doesn't degrade much as compared to the initial interactive model, then this implies that the interactive model doesn't actually learn complex cross-modal interactions and is just using an ensemble of unimodal components. Hence, we use the performance of the EMAP projected models as an intrinsic metric to evaluate the interactions learned by the models. Since, EMAP (Hessel and Lee, 2020) currently works only for classification based approaches, we evaluate this metric for the MMBT and M4 models using the Hateful Memes Dataset.

We list our results with and without EMAP in Table 5. It is evident from the results that there isn't a degradation in model performance. The EMAP projected models perform as good as the original interactive models, and this holds true for both MMBT and M4 models. This indicates that despite a gating mechanism being used in M4, it isn't learning cross modal interactions effectively. Hence, a more complex interaction module like Q-former might be more useful.

Model	Accuracy	ROC AUC
M4	0.5460	0.5761
M4 + EMAP	0.5440	0.5764
MMBT	0.52	0.5823
MMBT + EMAP	0.52	0.5823

Table 5: EMAP Scores of the M4 and MMBT models on Hateful Memes validation set

Overall, for both of our tasks, the intrinsic metrics indicate that models struggle to effectively capture cross-modal interactions. Further, high values obtained by InstructBLIP suggests that a Q-former based fusing module helps to better facilitate the interactions between the modalities as compared to using linear projection layers/gating mechanisms. The intrinsic metrics for the InstructBLIPFineTuned model show an increase in the ITC score, but drops in the ITM and Hit@K metrics suggest that the fusions learned in the fine-tuned model

VIT + GPT2: Meme poster is trying to convey that they don't need to buy Twitter to live broadcasts, they just need a dab pen

CLIP + GPT2: Poster is trying to help a guy named James that really wants to fight this girl in her own home.

LLaMA-7b: The meme poster is talking about an episode of The Simpsons. This episode has a character called Kid Pushy

InstructBLIP: the simpsons predicted it yet again

InstructBLIP + RAG(k=5): The meme poster is trying to convey that the Simpsons predicted it yet again.

InstructBLIPFineTuned: The person is a politician.

Ground Truth: The Simpsons was correct about it's use of Trump and Greta Thurnberg.

Figure 5: External Knowledge Failure Case

My nose at nights when I try to sleep:

VIT + GPT2: Meme poster is trying to convey that when training you must protect yourself at all cost

CLIP + GPT2: Meme poster wants to make sure his poster only lasts for about 15 minutes, but that doesn't seem to be working out

LLaMA-7b: The meme poster is trying to say that he used to have a problem with his nose which made breathing at nights

InstructBLIP: my nose at night when i try to sleep

InstructBLIP + RAG(k=5): The meme poster is trying to convey the idea that a person's nose can be a source of embarrassment, discomfort.

InstructBLIPFineTuned: This is a very old building. this is a cave

Ground Truth: Meme poster is trying to convey that their one nostril gets clogged up every time they try to sleep.

Figure 6: Multi-hop reasoning Failure Case

are not helpful.

7.2 Qualitative Analysis and Examples

We evaluated the generations of the ViT + GPT2, CLIP + GPT2, LLaMA, InstructBLIP, InstructBLIP + RAG, and InstructBLIPFineTuned on the Memecap test set. We present some of the failure samples in Figures 5 and 6. Meme caption generation is a very difficult task as the generated captions have to reflect the reasoning of the models. Further, from our qualitative analysis we find that there are numerous essential skills that the models should have in order to correctly understand the meaning of the memes. Some of those skills are listed as follow, and we present relevant failure cases of our baselines and proposed approaches.

1. **External Knowledge** is essential for understanding the memes. Even when the models can correctly identify the objects, external knowledge provides more information about what the objects metaphorically mean, and also helps to understand the appropriate relationship between the objects. In figure 5, ViT + GPT2 doesn't generate a relevant caption and fails to detect the appropriate objects in the meme. CLIP + GPT2 is able to identify that there are men and women in the meme, but fails to recognise the relationship between them. Since the prompt to all models contains some information (e.g., title) about *Simpsons*, LLaMA is able to generate response relevant to the title. However, it still fails to generate correct caption as it is unimodal. Finally, InstructBLIP generates the title as it is and doesn't add any new information. InstructBLIP + RAG is able to generate a good response as it focuses on the textual modality more than InstructBLIP. However, semantically the generations are similar to InstructBLIP. Upon further evaluation, we found that in our curated knowledge base, we don't have descriptions related to either of the characters in the meme. Since InstructBLIPFineTuned is instruction tuned to predict the inference from the visual cues, it is able to reason that someone in the image is a politician. However, since it does not have access to external knowledge, it fails on this case. Overall, all models fail to recognise the characters: *Trump* and *Greta Thunberg* and their use in the show *Simpsons*, due to the lack of external knowledge.

2. **Multi-step Reasoning** is essential for understanding memes as the objects in the memes are often used as metaphors. Therefore, in addition to reason about the alignment, models need to reason about the metaphorical use of the objects. For example, in Figure 6, the *tunnels* metaphorically refer to *nostrils*, and the models must be good at multi-step reasoning to figure this out. The generations of the models indicate that none of the models are able to properly reason about the metaphorical reference, and hence, generate inappropriate captions. However, with the help of retrieved captions, InstructBLIP + RAG is able to generate more relevant captions. InstructBLIPFineTuned also struggles and responses with its inference on the visual cues.

Majority of our evaluations reflect that ViT + GPT2 hallucinates and generates irrelevant captions. By using CLIP instead of ViT, the generations are better, but still the error cases show that the model struggles to use any of the aforementioned skills. LLaMA, although being only a Langueg Model, is good at generating captions when provided with the OCR text from the image. Moreover, since it has been trained on a lot of data, it is better able to use the knowledge as compared

Figure 7: Q: Is the meme hateful? **InstructBLIP:** Yes, **InstructBLIP + RAG:** No, **Ground Truth:** Yes

to GPT2. Finally, InstructBLIP always generates the text in the image, even when it is not provided with the OCR text. This implies that InstructBLIP is heavily using the vision features and not utilizing the language components much. InstructBLIP + RAG is able to more relevant captions, but the performance is highly limited by the quality of the knowledge base. Finally, InstructBLIPFineTuned is instruction tuned on the inference generationt task, hence, it always generates its inference about the image rather than connecting the inference to the metaphors used.

7.2.1 Classification Task

In our qualitative assessment of the hateful meme classification task (Figure 8), we employed five different models - two unimodal (TextBERT and Resnet) and three multimodal (M4, MMBT, and VisualBERT). While all models demonstrated competency on certain examples, there were notable instances where their performance suffered, revealing limitations that merit further examination.

One key challenge observed across models was the ability to comprehend nuanced context and subtext. Hateful rhetoric often employs dogwhistles and implicit biases that require deep understanding of cultural contexts. Models struggled with examples containing these nuances, likely due to their training data lacking sufficient representation of such examples. For instance, final meme in the dataset, depicting a man seemingly content with a beer in hand, juxtaposed with the text *we broke up but, she said we could still be cousins*. It is evident that all models failed to identify this as hateful. This scenario underscores the potential benefits of enriching the training data and incorporating external knowledge bases to enhance the comprehension of complex memes.

Analysing the fourth meme, highly offensive and targeting a specific community, posed a challenge for most models, yet TextBERT correctly predicted its classification. This success can be attributed to TextBERT's fine-tuning on hateful memes, likely enabling it to recognize and categorize such instances based on its exposure to similar language and contexts during training, including terms like "mental illness." In the fifth and sixth meme, all four models except TextBERT predicted it to be hateful, which suggests that image embeddings played a crucial role in their success. This highlights the importance of multimodal approaches, where the integration of visual features enables models to capture nuanced visual cues and recognize racially charged imagery, contributing to improved classification accuracy.

For the second meme, MMBT and M4 struggled to perform well as compared to TextBERT and VisualBERT, as the hateful content was primarily in the text, while the image appeared benign. Both MMBT and M4 fusion mechanisms might not effectively capture the derogatory language, which may have diluted the discriminatory signal. In contrast, TextBERT's focus on text and VisualBERT's parallel processing of text and image likely enabled them to accurately detect the hateful language. The unimodal ResNet model, focusing solely on images, avoided being misled by the relatively innocuous image content. This suggests a potential weakness in MMBT and M4's embedding projection approach for handling unbalanced multimodal inputs.

In the seventh and eighth memes, it's noteworthy that while the textual content remains consistent, the visual elements differ. What's intriguing is how the M4 model's prediction shifts in response to these visual alterations. This sheds light on the fusion dynamics within the MMBT and M4 models. MMBT employs a concatenation of textual and visual embeddings derived from BERT and ResNet. Consequently, both memes yield identical predictions from MMBT, as expected. Conversely, M4 utilizes distinct embeddings and employs a gated fusion mechanism, likely enabling it to more effectively capture the multimodal nuances of the memes.

Quantitative results show that InstructBLIP achieves the best performance, and with RAG the performance deteriorates. To further explore the reasons behind the deterioration, we performed qualitative analysis on the generations of InstructBLIP and InstructBLIP + RAG. We find that, the samples that InstructBLIP was predicting correctly, were wronged by InstructBLIP + RAG due to the

retrieved samples which were not helpful. See Figure 7 for an example. Here, the retrieved captions are non-hateful in nature and irrelevant to the meme which totally misleads the model towards predicting the non-hateful label.

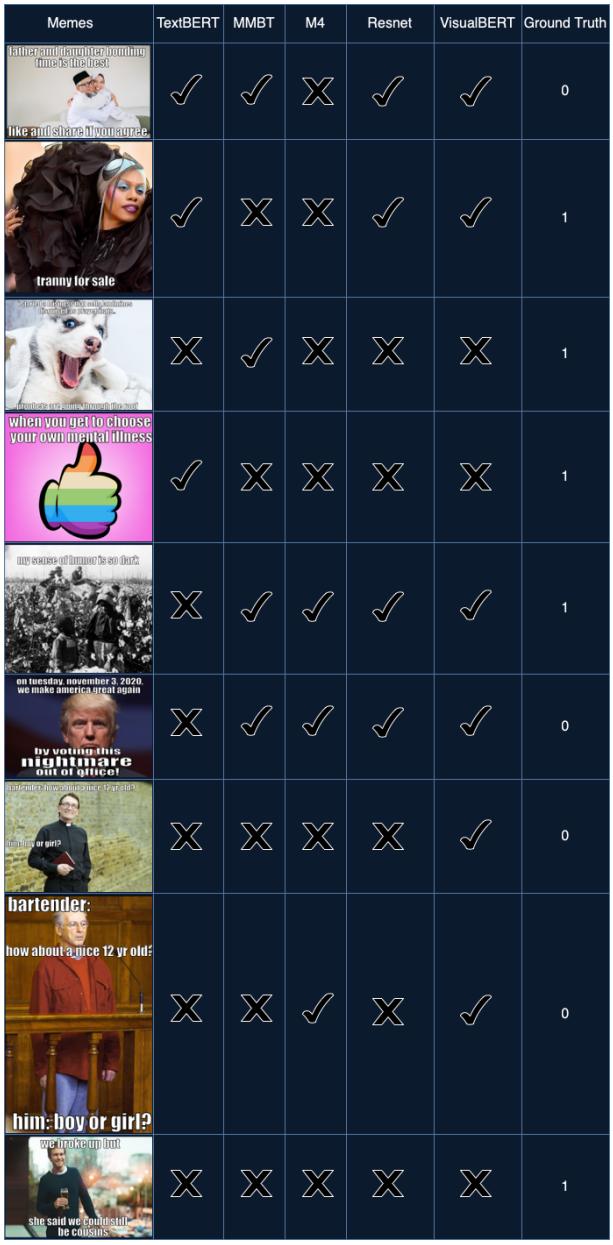

Figure 8: Results of models on selected memes for hateful meme classification task. The check mark indicates that the corresponding model got the right prediction and the cross denotes otherwise. the ground truth is 0 for non-hateful memes and 1 for hateful.

8 Future work and Limitations

While our proposed approaches of using Retrieval Augmented Generation (RAG) and fine-tuning on abductive reasoning datasets show promise, there are several limitations and avenues for future work.

The RAG approach suffers from the following limitations:

1. **Quality of Retrieved Knowledge**: The performance of the RAG approach crucially depends on the quality and relevance of the retrieved knowledge snippets. In our work, we used a basic CLIP-based retrieval mechanism over a limited knowledge base. More sophisticated techniques are needed to retrieve precise and contextually relevant knowledge for memes, which often require nuanced cultural understanding.

2. **Static Knowledge Base**: Our current knowledge base is static and cannot account for rapidly evolving meme trends and references to current events. Continuous updating of the knowledge source is essential to maintain relevance over time.

3. **Knowledge Base Limitations**: The knowledge sources we used like TV captions and manual annotations have limited coverage and may not generalize well across diverse meme styles, formats, and topics. More extensive and structured multimodal knowledge bases are required. This problem also includes the issue of inherent biases present in a dataset and challenge of debiasing them.

Similarly, fine-tuning a model to learn abductive reasoning is also not without its challenges.

1. As pointed out in the qualitative results, by fine-tuning on the SHERLOCK dataset the models learn to infer properly using visual cues. However, for the memecap dataset, we see deterioration in performance as the model is not fine-tuned to also utilize the text modality. For future work, we will also experiment with adding the clues in the SHERLOCK dataset in the textural prompt to enhance the interaction.

2. We performed full fine-tuning of the Q-former. In future, we would also like to explore the LORA fine-tuning of the fusion module.

Our work opens several promising avenues for future research that can build upon the foundations of this project:

1. **Early Multimodal Fusion with Knowledge**: Current fusion techniques may be insufficient for integrating external knowledge. Exploring early, tighter fusion of multimodal representations and knowledge snippets could lead to more effective knowledge grounding.

2. **Leveraging Structured Knowledge**: Instead of unstructured text snippets, using structured knowledge bases like knowledge graphs may allow richer encoding of semantic relationships and reasoning paths beneficial for memes.

3. **Continuous Learning**: Developing techniques to continuously expand and adapt meme understanding models as new trends, events, and cultural shifts emerge over time.

4. **Robustness to Adversarial Inputs**: Memes can often involve adversarial editing to convey offensive meanings, affect caption generation, and more. Improving model and representation robustness against such adversarial attacks is crucial.

The ethical considerations surrounding potential biases, misinformation, and dual-use risks of meme analysis models (more in section 9) also necessitate careful analysis and mitigation strategies in future research. Overall, meme understanding epitomizes many open challenges in multimodal AI, offering rich opportunities for further exploration and innovation.

9 Ethical Concerns and Considerations (unintentional, malicious, and dual-use)

Dealing with memes is inherently a sensitive task given how polarizing memes can be. There are ethical concerns surrounding almost every aspect of analyzing hateful memes. Especially when it comes to incorporating external knowledge. While such knowledge can improve accuracy, it's crucial to consider the source and potential bias. It is possible that the knowledge base itself can be biased towards a certain race, ethnicity or religion. These models might inherit societal prejudices from the external data, leading to misinterpretations of memes that hinge on cultural nuances or satire. Moreover, it can be debated if using ethnicity, race, etc. for the task is even ethical in this context. This calls for greater explainability and interpretability of these models.

Multimodal models on this domain can exhibit unintentional biases against certain groups of people and may even fail to detect memes that are actually harmful. Furthermore, with better cross-modal interaction and enhanced meme understanding that these models may display, people with malicious intent can use these representations to generate more polarizing and hate content online. Increasingly dangerous meme captions for any given image is one example. Malicious actors could also leverage these generators to create misleading memes that spread misinformation or propaganda. The models' ability to mimic existing trends could make these fakes particularly believable. Transparency and trust is going to be crucial in the deployment and productionizing of these models.

Moreover, memes are an inherently subjective medium, with interpretations varying substantially across individuals, making it extremely difficult to establish definitive benchmarks or thresholds for determining potentially harmful content. This subjectivity is further compounded by the reliance on manually annotated datasets for training, where inter-annotator agreement can fluctuate considerably across different datasets, leading to inconsistencies in the ground truth labels used to train these models.

While mitigating the proliferation of hateful memes on social media platforms is an important task, the development and deployment of such models must be approached with utmost care and cognizance of the profound ethical implications. If deployed for real-time meme classification, these models could potentially impinge upon fundamental rights such as free speech and expression, necessitating a multi-stakeholder approach involving researchers, policymakers, and the broader public.

To address these myriad ethical concerns, the development of explainable and interpretable models is imperative, fostering transparency and accountability in their decision-making processes. Robust mechanisms for model evaluation, validation, and auditing must be established to ensure that these models are not perpetuating harmful biases or being misused for malicious purposes.

In essence, the development and deployment of multimodal models for meme understanding and classification tasks necessitate a holistic understanding of the ethical implications and potential societal impacts. Striking the delicate balance between mitigating the spread of harmful content and

preserving fundamental rights will require a concerted, multi-disciplinary effort involving diverse stakeholders and an unwavering commitment to ethical principles.

10 Team member contributions

Member 1 contributed to prior work review, unimodal and multimodal baselines and ablations for hateful memes, future work.

Member 2

1. Model Proposal: Proposed both models and implemented them.
2. Curating Knowledge Base for RAG: TVC Caption Dataset
3. Literature Review: Reviewd all the literature related to reasoning and RAG based approaches.
4. Training/Finetuining: Fine-tuned the InstructBLIPFineTuned model on the SHERLOCK datset and also implemnented the RAG approach: both retrieval through clip and inference using instructblip
5. Running experiments, instrinsic metrics and error analysis on baselines and proposed models
6. Writing: Wrote the Abstract, Introduction, Proposed model, Task setup, Results and Analysis of Proposed models solely

Member 3 contributed to prior work review, relevant techniques, dataset (knowledge base creation for RAG), baselines and multimodal baselines for classification task, future work and limitations, ethical concerns.

Member 4 Worked on prior work review and curating the knowledge base for RAG: Decoding the Underlying Meaning of Multimodal Hateful Memes Dataset. Implemented the unimodal,multimodal baselines for hateful memes classification. Worked on ethical concerns and future work.

References

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kenton Lee Kristina Toutanova Devlin J., Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Samar Elbedwehy, T Medhat, Taher Hamza, and Mohammed F Alrahmawy. 2022. Efficient image captioning based on vision transformer models. *Computers, Materials & Continua*, 73(1).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Biagio Grasso, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. 2024. Kermit: Knowledge-empowered model in harmful meme detection. *Information Fusion*, 106:102269.

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

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes.

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

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from

the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020a. Supervised multimodal bitransformers for classifying images and text.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

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.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

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

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*.

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

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

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

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023a. Smallcap: Lightweight image captioning prompted with retrieval augmentation. *CVPR*.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023b. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2840–2849.

Rita Parada Ramos, Bruno Martins, and Desmond Elliott. 2023c. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. *ArXiv*, abs/2305.19821.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Shardul Suryawanshi, Mihael Arcan, Suzanne Little, and Paul Buitelaar. 2023. Multimodal offensive meme classification with natural language inference. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 134–145.

Chow Terence. 2021. Memotion multi-modal. https://github.com/terenceylchow124/Meme-MultiModal.

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

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

Haofei Yu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023. Mmoe: Mixture of multimodal interaction experts. *arXiv preprint arXiv:2311.09580*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6.

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

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution.