

Singapore Management University

## Institutional Knowledge at Singapore Management University

---

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

---

11-2023

### Pro-Cap: Leveraging a frozen vision-language model for hateful meme detection

Rui CAO

Singapore Management University, ruicao.2020@phdcs.smu.edu.sg

Ming Shan HEE

Adriel KUEK

Wen Haw CHONG

Singapore Management University, whchong.2013@phdis.smu.edu.sg

Roy Ka-Wei LEE

*See next page for additional authors*

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)



Part of the [Databases and Information Systems Commons](#), [Graphic Communications Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

---

#### Citation

CAO, Rui; HEE, Ming Shan; KUEK, Adriel; CHONG, Wen Haw; LEE, Roy Ka-Wei; and JIANG, Jing. Pro-Cap: Leveraging a frozen vision-language model for hateful meme detection. (2023). *MM '23: Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, October 29 - November 3*. 5244-5252. Available at: [https://ink.library.smu.edu.sg/sis\\_research/8477](https://ink.library.smu.edu.sg/sis_research/8477)

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [cherylids@smu.edu.sg](mailto:cherylids@smu.edu.sg).

---

**Author**

Rui CAO, Ming Shan HEE, Adriel KUEK, Wen Haw CHONG, Roy Ka-Wei LEE, and Jing JIANG



# Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection

Rui Cao  
ruicao.2020@phdcs.smu.edu.sg  
Singapore Management University  
Singapore, Singapore

Ming Shan Hee  
mingshan\_hee@mymail.sutd.edu.sg  
Singapore University of Design and  
Technology  
Singapore, Singapore

Adriel Kuek  
adriekuek@gmail.com  
DSO National Laboratories  
Singapore, Singapore

Wen-Haw Chong  
whchong.2013@phdis.smu.edu.sg  
Singapore Management University  
Singapore, Singapore

Roy Ka-Wei Lee  
roy\_lee@sutd.edu.sg  
Singapore University of Design and  
Technology  
Singapore, Singapore

Jing Jiang  
jingjiang@smu.edu.sg  
Singapore Management University  
Singapore, Singapore

## ABSTRACT

Hateful meme detection is a challenging multimodal task that requires comprehension of both vision and language, as well as cross-modal interactions. Recent studies have tried to fine-tune pre-trained vision-language models (PVLMs) for this task. However, with increasing model sizes, it becomes important to leverage powerful PVLMs more efficiently, rather than simply fine-tuning them. Recently, researchers have attempted to convert meme images into textual captions and prompt language models for predictions. This approach has shown good performance but suffers from non-informative image captions. Considering the two factors mentioned above, we propose a probing-based captioning approach to leverage PVLMs in a zero-shot visual question answering (VQA) manner. Specifically, we prompt a frozen PVLm by asking hateful content-related questions and use the answers as image captions (which we call Pro-Cap), so that the captions contain information critical for hateful content detection. The good performance of models with Pro-Cap on three benchmarks validates the effectiveness and generalization of the proposed method.<sup>1</sup>

## CCS CONCEPTS

• Computing methodologies → Natural language processing; Computer vision representations.

## KEYWORDS

memes, multimodal, semantic extraction

### ACM Reference Format:

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada.

<sup>1</sup>Code is available at: <https://github.com/Social-AI-Studio/Pro-Cap>



This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada  
© 2023 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0108-5/23/10.  
<https://doi.org/10.1145/3581783.3612498>

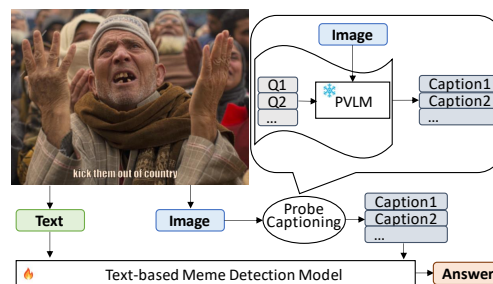


Figure 1: The proposed probe-captioning approach. We prompt frozen pre-trained vision-language models via visual question answering to generate hateful content centric image captions.

ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3612498>

**Disclaimer:** This paper contains violence and discriminatory content that may be disturbing to some readers.

## 1 INTRODUCTION

Memes, which combine images with short texts, are a popular form of communication in online social media. Internet memes are often intended to express humor or satire. However, they are increasingly being exploited to spread hateful content across online platforms. Hateful memes attack individuals or communities based on their identities such as race, gender, or religion [5, 8, 12, 27]. The propagation of hateful memes can lead to discord online and may potentially result in hate crimes. Therefore, it is urgent to develop accurate hateful meme detection methods.

The task of hateful meme detection is challenging due to the multimodal nature of memes. Detection involves not only comprehending both the images and the texts but also understanding how these two modalities interact. Previous work [14, 28, 35, 36] learns cross-modal interactions from scratch using hateful meme detection datasets. However, it may be difficult for models to learn complicated multimodal interactions with the limited amount of

**Table 1: Impact on detection performances on the FHM dataset [12] from image captions. (w/o) denotes models without additional entity and demographic information.**

Model	Performance	
	AUC	Acc.
PromptHate (w/o)	76.76	67.28
PromptHate	81.45	72.98
VisualBERT (w/o)	68.71	61.48
VisualBERT	72.56	68.24
ViLBERT (w/o)	73.05	64.70
ViLBERT	75.72	68.24

data available from these datasets. With the development of Pre-trained Vision-Language Models (PVLMs) such as VisualBERT [18] and ViLBERT [23], recent work leverage these powerful PVLMs to facilitate the hateful meme detection task. A common approach is to fine-tune PVLMs with task-specific data [9, 20, 26, 34, 37]. However, it is less feasible to fine-tune the larger models such as BLIP-2 [15] and Flamingo [1] on meme detection because there are billions of trainable parameters. Therefore, computationally feasible solutions other than direct fine-tuning are needed to leverage large PVLMs in facilitating hateful meme detection.

Different from the approach above using PVLMs, PromptHate[2] is a recently proposed model that converts the multimodal meme detection task into a unimodal masked language modeling task. It first generates meme image captions with an off-the-shelf image caption generator, ClipCap[25]. By converting all input information into text, it can prompt a pre-trained language model along with two demonstrative examples to predict whether or not the input is hateful by leveraging the rich background knowledge in the language model. Although PromptHate achieves state-of-the-art performance, it is significantly affected by the quality of image captions, as shown in Table 1. Image captions that are merely generic descriptions of images may omit crucial details [14, 37], such as the race and gender of people, which are essential for hateful content detection. But with additional image tags, such as entities found in the images and demographic information about the people in the images, the same model can be significantly improved, as shown in Table 1. However, generating these additional image tags is laborious and costly. For instance, entity extraction is usually conducted with the Google Vision Web Entity Detection API<sup>2</sup>, which is a paid service. Ideally, we would like to find a more affordable way to obtain entity and demographic information from the images that is critical for hateful content detection.

Both above-mentioned approaches (i.e., one using PVLMs and the other converting the task to a unimodal task) have their pros and cons. In this paper, we combine the ideas from these two approaches and design a hateful meme detection method that leverages the power of a frozen PVLm to complement the unimodal approach of PromptHate. Specifically, we use a set of “probing” questions to query a PVLm (BLIP-2 [15] in our experiments) for information related to common vulnerable targets in hateful content. The answers

obtained from the probing questions will be treated as image captions (denoted as **Pro-Cap**) and used as input to a trainable hateful meme detection model. Figure 1 illustrates the overall workflow of the method. We refer to the step of using probing questions to generate the captions as *probing-based captioning*.

Our proposed method fills existing research gaps by: 1) Leverage a PVLm without any adaptation or fine-tuning, thereby reducing computational cost; 2) Instead of explicitly obtaining additional image tags with costly APIs, we utilize the frozen PVLm to generate captions that contain information useful for hateful meme detection. To the best of our knowledge, this is the first work that to leverage PVLms in a zero-shot manner through question answering to assist in the hateful meme detection task. To further validate our method, we test the effect of the generated Pro-Cap on both PromptHate[2] and a BERT-based[4] hateful meme detection model.

Based on the experimental results, we observe that PromptHate with Pro-Cap (denoted as Pro-CapPromptHate) significantly surpasses the original PromptHate without additional image tags (i.e., about 4, 6, and 3 percentage points of absolute performance improvement on FHM [12], MAMI [5], and HarM [28] respectively). Pro-CapPromptHate also achieves comparable results with PromptHate with additional image tags, indicating that probing-based captioning can be a more affordable way of obtaining image entities or demographic information. Case studies further show that Pro-Cap offers essential image details for hateful content detection, enhancing the explainability of models to some extent. Meanwhile, Pro-CapBERT clearly surpasses multimodal BERT-based models of similar sizes (i.e., about 7 percentage points of absolute improvement with VisualBERT on FHM [12]), proving the generalization of the probing-based captioning method.

## 2 RELATED WORK

*Memes*, typically intended to be humorous or sarcastic, are increasingly being exploited for the proliferation of hateful content, leading to the challenging task of online hateful meme detection [5, 12, 27]. To combat the spread of hateful memes, one line of work regards the hateful meme detection as a multimodal classification task. Researchers have applied pre-trained vision-language models (PVLms) and fine-tune them based on meme detection data [20, 26, 34, 37]. To improve performance, some have tried model ensembling [20, 26, 34]. Another line of work considers combining pre-trained models (e.g., BERT [4] and CLIP [29]) with task-specific model architectures and tunes them end-to-end [13, 14, 28]. Recently, authors in [2] have tried converting all meme information into text and prompting language models to better leverage the contextual background knowledge present in language models. This approach achieves the state-of-the-art results on two hateful meme detection benchmarks. However, it adopts a generic method for describing the image through image captioning, often ignoring important factors necessary for hateful meme detection. In this work, we seek to address this issue through probe-based captioning by prompting pre-trained vision-language models with hateful content-centric questions in a zero-shot VQA manner.

<sup>2</sup><https://cloud.google.com/vision/docs/detecting-web>

### 3 PRELIMINARY

We formally define our task and briefly review the use of pre-trained vision-language models (PVLMs) for zero-shot visual question answering (VQA). At the end of the section, we provide a brief introduction to the specific PVLm utilized in our work.

Given a meme image  $\mathcal{I}$  and a piece of accompanying meme text  $\mathcal{T}$ , the model predicts whether the meme is hateful or not. Specifically, the model predicts scores  $\mathbf{s} \in \mathbb{R}^2$  over the label space, where  $s_0$  is a score indicating how likely the meme is *non-hateful*, whereas  $s_1$  is a score for the meme being *hateful*. If  $s_0 > s_1$ , the model classifies the meme as non-hateful; otherwise, the meme is classified as hateful. Our proposed method (to be presented in detail in Section 4) uses zero-shot VQA to generate relevant captions to assist with hateful meme detection. To perform zero-shot VQA, we assume that there is a PVLm capable of processing an image and a textual prompt formatted as *Question: [QUESTION] Answer:*, where [QUESTION] is a placeholder for the question. The PVLm then generates a sequence of tokens as the answer to the question. For example, given an image showing an Asian woman and the prompt *Question: What is the race of the person in the image? Answer:*, the PVLm may generate the answer *Asian*.

In this work, we use the recently released BLIP-2 model [15] as the PVLm, as it has demonstrated good performance in zero-shot VQA. The BLIP-2 model is composed of a frozen pre-trained image encoder, a frozen pre-trained language model, and a lightweight Querying Transformer, which is responsible for bridging the modality gap. It is worth noting that the BLIP-2 model can be replaced with any other PVLm that is capable of zero-shot VQA.

## 4 PROPOSED METHOD

### 4.1 Overview

Recall that the key idea of our method is to elicit image details that are critical for hateful content detection, such as the gender and race of the people in the image. Because these details are not always included in automatically generated image captions, we propose relying on VQA to obtain such critical information, where the questions are carefully curated to elicit demographic and other relevant information. We opt to use zero-shot VQA because (1) for the intended type of questions, we do not have any VQA training data to train our own model, and (2) recent work has demonstrated promising performance of zero-shot VQA.

Specifically, we prompt the PVLm with  $K$  *probing questions* and regard the set of  $K$  answers from the PVLm as image captions, which we refer to as **Pro-Cap**. We then combine the original text  $\mathcal{T}$  with Pro-Cap as input to a hateful meme detection model. We experiment with two alternative hateful meme detection models: one based on BERT encoding, and the other based on PromptHate, a recently proposed prompting-based hateful meme detection model.

In the rest of this section, we first present the details of how we design our VQA questions to elicit the most critical details of an image for hateful meme detection. We then explain how the generated Pro-Cap is used by two alternative hateful meme detection models.

### 4.2 Design of VQA Questions

We leverage PVLms for zero-shot VQA to generate Pro-Cap as image captions. We want Pro-Cap to provide not only a general description of the image but also details critical for hateful meme detection. To obtain a general caption of the image, we design the first probing question to inquire about the generic content of the image, as shown in Table 2. However, such generic captions may be insufficient for hateful meme detection as hateful content usually targets persons or groups with specific characteristics, such as race, gender, or religion [5, 12]. Additionally, previous studies have shown that augmenting image representations with entities found in the image or demographic information of people in the image significantly aids hateful meme detection [14, 37]. Such details may be missing in generic image captions. Therefore, we design additional questions that aim to bring out information central to hateful content. This aligns the generated image captions more closely with the goal of hateful meme detection. Specifically, the high-level idea is to ask questions about common vulnerable targets of hateful content. Inspired by [24], which categorizes the targets of hateful memes into *Religion, Race, Gender, Nationality, and Disability*, we ask questions about these five types of targets. For example, to generate image captions that indicate the race of the people in an image, we can ask the following question: *what is the race of persons in the image?* We list the five questions designed for these five types of targets in Table 2. Additionally, we observe that some animals, such as pigs, are often depicted in hateful memes, frequently as a means to annoy Muslims. With this consideration, we also design a question asking about the presence of animals in the image.

In [3], the author claimed that PVLms may hallucinate non-existent objects. For example, even when there is nobody in an image, PVLms may generate an answer about race in response to the question *what is the race of the person in the image?*. To prevent such misleading information, we use two validation questions. Specifically, we inquire about the existence of persons and animals. Only when the PVLm responds that a person or an animal exists will we include in the Pro-Cap the answers to those person-related or animal-related questions. For instance, if the answer to the question validating the existence of people indicates that nobody is present, we will ignore all answers from questions asking about *religion, race, gender, nationality, and disability*.

We use  $C$  to represent the concatenation of the answers to the probing questions that are finally included as part of the Pro-Cap based on the validation results. We will then concatenate  $\mathcal{T}$  and  $C$  together as input to a purely text-based hateful meme classification model, as shown at the bottom of Figure 1.

### 4.3 BERT-based Detection Model

We now introduce the first of the two alternative hateful meme classification models, which is based on BERT [4]. We first feed the concatenation of the meme text  $\mathcal{T}$  and the Pro-Cap  $C$  into the BERT model to generate a vector  $\mathbf{r} \in \mathbb{R}^d$ :

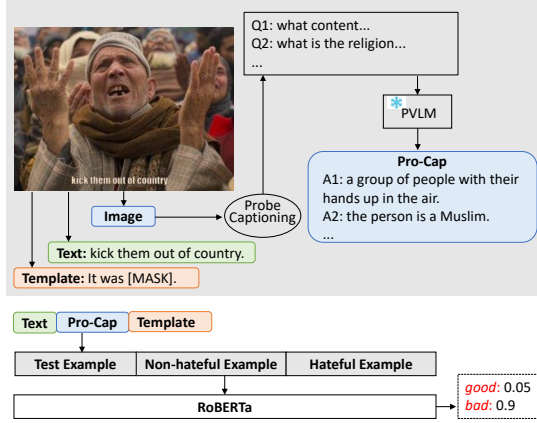
$$\mathbf{r} = \text{BERT}([\mathcal{T}, C]), \quad (1)$$

where  $[\cdot, \cdot]$  represents concatenation. Next, we feed the sentence representation  $\mathbf{r}$  into a linear layer for hateful meme classification:

$$\mathbf{s} = \text{Sigmoid}(\mathbf{W}^T \mathbf{r} + \mathbf{b}), \quad (2)$$

**Table 2: Details of questions prompting PVLMs. The first block of the question asks about the content of the image; questions in the second block ask about commonly seen vulnerable targets in hateful contents; the last block questions validate the existence of persons and animals.**

Focus	Questions
Content	what is shown in the image?
Race	What is the race of the person in the image?
Gender	What is the gender of the person in the image?
Religion	What is the religion of the person in the image?
Nationality	Which country does the person in the image come from?
Disability	Are there disabled people in the image?
Animal	What animal is in the image?
Val Person	Is there a person in the image?
Val Animal	Is there an animal in the image?



**Figure 2: An overview of the PromptHate model and how pro-cap is used in PromptHate.**

where  $\mathbf{W} \in \mathbb{R}^{d \times 2}$  and  $\mathbf{b}^2$  are learnable parameters.

#### 4.4 PromptHate for Hateful Meme Detection

Next, we introduce the second hateful meme classification model, PromptHate [2], which employs a prompt-based method to classify memes. PromptHate was developed to better leverage contextual background knowledge by prompting language models. Given a test meme, PromptHate first uses an image captioning model to obtain generic image captions. It then concatenates the meme text, the image captions, and a prompt template into  $\mathcal{S}$ : *It was [MASK].*, to prompt a language model (LM) to predict whether the meme is hateful. Specifically, it compares the probability of the language model predicting [MASK] to be a positive word (e.g., *good*) given

the context, versus the probability of predicting a negative word (e.g., *bad*). The approach also includes one positive and one negative example in the context, and [MASK] will be replaced by their respective label words. An overview of PromptHate is shown in Figure 2. For further details, please refer to [2].

In [2], PromptHate utilizes ClipCap [25] to generate image captions. In this work, we replace this with Pro-Cap  $\mathcal{C}$ . We then represent every meme  $\mathcal{O}$  as  $\mathcal{O} = [\mathcal{T}, \mathcal{C}, \mathcal{S}]$ . With these inputs, the language models (LMs), for instance, RoBERTa [21], generate confidence scores for the masked word over their vocabulary space,  $\mathcal{V}$ :

$$\mathbf{p} = \text{Sigmoid}(\text{LM}([\mathcal{O}_{\text{test}}, \mathcal{O}_{\text{non-hate}}, \mathcal{O}_{\text{hate}}])), \quad (3)$$

where  $\mathbf{p} \in \mathbb{R}^{|\mathcal{V}|}$ . We extract the score for the label words as the prediction:

$$s_0 = p_i, \quad \mathcal{V}_i = \mathcal{W}_{\text{pos}}, \quad (4)$$

$$s_1 = p_j, \quad \mathcal{V}_j = \mathcal{W}_{\text{neg}}. \quad (5)$$

#### 4.5 Model Training and Prediction

We denote the ground-truth label of a meme as  $\hat{y} \in \mathbb{R}^2$ . If the meme is annotated as *non-hateful*,  $\hat{y}_0$  will be 1 while  $\hat{y}_1$  will be 0, otherwise,  $\hat{y} = [0, 1]$ . The binary cross-entropy loss is applied for model training:

$$\text{Loss} = -(\hat{y}_0 * \log(s_0) + \hat{y}_1 * \log(s_1)). \quad (6)$$

For model prediction, if  $s_0 > s_1$ , the meme will be predicted as non-hateful, otherwise, hateful.

## 5 EXPERIMENT

In this section, we first introduce our evaluation datasets, metrics and implementation details. Next, we introduce the baselines for comparison. Finally, we conduct qualitative analysis with case studies and error analysis to better understand the advantages and limitations of our method.

### 5.1 Experiment Settings

**Evaluation Datasets.** We test our proposed method on benchmarks for hateful meme detection. We evaluate our method on three datasets to better illustrate the generalization and stability of our approach. Table 3 presents the statistics of these datasets.

The *Facebook Hateful Meme* dataset (FHM) [12] was constructed by Facebook. It contains synthetic memes with added confounders such that unimodal information is insufficient for detection and deep multimodal reasoning is required. The FHM dataset contains hateful memes targeting various vulnerable groups in categories including *Religion*, *Race*, *Gender*, *Nationality*, and *Disability*. As the labels of the test split of FHM are not available, we perform evaluation on its *dev-seen* split.

Different from FHM, the *Multimedia Automatic Misogyny Identification* (MAMI) dataset focuses on a particular type of hateful memes, namely, those targeting women. Performance on MAMI therefore reflects the capability of hateful meme detection methods for female victims.

To test our method’s generalization capability, we also consider a harmful meme detection dataset, **HarM** [27]. HarM contains

**Table 3: Statistical distributions of datasets used for evaluation.**

Datasets	Train		Test	
	#Hate.	#Non-hate.	#Hate.	#Non-hate.
FHM	3,019	5,481	247	253
HarM	1,064	1,949	124	230
MAMI	5,004	4,995	500	495

memes related to COVID-19, which are classified into three categories: *harmless*, *partially harmful*, and *very harmful*. We merge *partially harmful* and *harmful* into one category. Because hateful content is always regarded as harmful, we use this dataset to test the capability of generalization of our proposed method from hateful meme detection to harmful meme detection.

**Evaluation Metrics.** Hateful meme detection is a binary classification task. In addition to detection accuracy, we also compute the Area Under the Receiver Operating Characteristics curve (AU-CROC) used in prior work [2, 14, 20, 37]. We conduct experiments with **ten** random seeds and report the average performance and standard deviation. All models use the same set of random seeds.

**Implementation Details.** Given a meme image, we first detect the meme text with the open-source Easy-OCR tool<sup>3</sup> and then in-paint over the detected texts. To generate the answers to VQA questions, we prompt BLIP-2 [15], specifically the FlanT5<sub>XL</sub> version. We then insert the generated image captions into two text-based hateful meme detection models, i.e., the BERT-based model and the PromptHate model. For the BERT-based model, to avoid overfitting, we add a dropout rate of 0.4 to the classification layer. We use a learning rate of  $2e-5$  and a batch size of 64. For PromptHate, we train the model with a batch size of 16 and empirically set the learning rate to  $1.3e-5$  on FHM and  $1e-5$  on the other two datasets [6]. We optimize both models with the AdamW optimizer [22] and implement them in PyTorch. Due to space limit, we provide more details (i.e., computation costs and model sizes) in Appendix.

## 5.2 Baselines

We compare our method against both unimodal and multimodal models to demonstrate the effectiveness of the proposed method, where we regard models receiving information from one modality (i.e., the meme text or the meme image only) as unimodal models. Note that because Pro-Cap already contains image information, even if Pro-Cap is input into a unimodal BERT, the model is not considered to be unimodal.

For the unimodal models, we consider a text-only and an image-only model. For the text-only model, we fine-tune a pre-trained BERT model [4] based on the meme text only for meme classification, which we represent as **Text-BERT**. For the image-only model, we first extract object-level image features with an off-the-shelf feature extractor, Faster-RCNN [30], which is trained for object detection. We then perform average pooling over object features and feed the resulting vector into a classification layer. We use **Image-Region** to denote the image-only model.

<sup>3</sup><https://github.com/JaidedAI/EasyOCR>

For multimodal models, we categorize them into two groups: 1) fine-tuning generic multimodal models that are proposed to conduct different multimodal tasks; 2) models specifically designed for hateful meme detection. For the first type of multimodal models, we firstly consider the **MMBT-Region** model [11], which is a widely used multimodal baseline in hateful meme detection [2, 12, 28] and the model has not been pre-trained with multimodal data. Secondly, we consider several multimodal pre-trained models, such as VisualBERT [18] pre-trained on MS-COCO [19] (**VisualBERT COCO**) and ViLBERT pre-trained on Conceptual Captions [32] (**ViLBERT CC**). Some recently released powerful pre-trained models are also included such as the *Align before Fusion* model [17] (**ALBEF**) and the *Bootstrapping Language-Image Pre-training* model [16] (**BLIP**). For the second category of baselines which are designed for the meme detection task, we consider the models listed below. The **CLIP-BERT** model [28] leverages the CLIP model [29] to deal with noisy meme images, uses pre-trained BERT [4] for representing meme text, and fuses them with concatenation. The **MOMENTA** model [28] designed both local and global multimodal fusion mechanisms to exploit multimodal interactions for hateful meme detection. Note that the MOMENTA model is designed to leverage augmented image tags (the detected image entities). **DisMultiHate** [14] disentangles target information from memes as targets are essential for identifying hateful content. The **PromptHate** model [2] is what we discussed in Section 4.4.

## 5.3 Experiment Results

As discussed earlier, previous work has shown that additional image tags can enhance hateful meme detection. We therefore consider two settings for comparison: 1) without any augmented image tags; 2) with augmented image tags. We display the performance of models **without** augmented image tags in Table 4 and **with** augmented image tags in Table 5. The standard deviations ( $\pm$ ) of ten random seed runs are also reported, and the best results are highlighted in bold.

**Without augmented image tags:** We first compare Pro-CapBERT with unimodal and multimodal models that also utilize BERT as the text encoder (i.e., VisualBERT, ViLBERT, and MMBT-Region). Evidently, Text BERT, which utilizes only meme text, is substantially outperformed by Pro-CapBERT. This suggests that 1) visual signals are vital for hateful meme detection, and 2) the image captions obtained from the probing questions are informative.

Experiment results from multimodal pre-trained BERT-based models are presented in the second block of Table 4. Interestingly, Pro-CapBERT still has better performances in all three datasets, surpassing the most powerful multimodal pre-trained BERT-base model, ViLBERT, by over 4% on FHM and surpassing MMBT-Region by about 3% on HarM. This is despite the fact that BERT has less model parameters compared with these multimodal models (e.g., ViLBERT has 252.1M parameters while BERT only has about 110M parameters). Pro-CapBERT is still competitive against models specifically designed for hateful meme detection (i.e., models in the third block of Table 4). We provide experimental results of recently published multimodal pre-trained models (i.e., BLIP and ALBEF) in the fourth block. By comparing the simple Pro-CapBERT with these models, we observe that Pro-CapBERT gives comparable results.

**Table 4: Model comparison without any augmented image tags.**

Dataset Model	FHM		MAMI		HarM	
	AUC.	Acc.	AUC.	Acc.	AUC.	Acc.
Text BERT	66.10 $\pm$ 0.55	57.12 $\pm$ 0.49	74.48 $\pm$ 0.60	67.37 $\pm$ 0.57	81.39 $\pm$ 0.91	75.68 $\pm$ 1.59
Image-Region	56.69 $\pm$ 1.05	52.34 $\pm$ 1.39	70.20 $\pm$ 0.63	64.18 $\pm$ 0.81	76.46 $\pm$ 0.47	73.05 $\pm$ 1.80
VisualBERT COCO	68.71 $\pm$ 1.02	61.48 $\pm$ 1.19	78.71 $\pm$ 0.59	71.06 $\pm$ 0.94	80.46 $\pm$ 1.04	75.31 $\pm$ 1.44
ViLBERT CC	73.05 $\pm$ 0.62	64.70 $\pm$ 1.12	77.71 $\pm$ 1.20	69.48 $\pm$ 1.00	84.11 $\pm$ 0.88	78.70 $\pm$ 1.17
MMBT-Region	72.86 $\pm$ 0.64	65.06 $\pm$ 1.76	79.17 $\pm$ 0.91	70.46 $\pm$ 0.76	85.48 $\pm$ 0.75	79.83 $\pm$ 2.00
CLIP-BERT	66.97 $\pm$ 0.34	58.28 $\pm$ 0.63	77.66 $\pm$ 0.64	68.44 $\pm$ 1.07	82.63 $\pm$ 3.83	80.48 $\pm$ 1.95
DisMultiHate	69.11 $\pm$ 0.84	62.42 $\pm$ 0.72	78.21 $\pm$ 0.61	70.58 $\pm$ 1.13	83.69 $\pm$ 1.33	78.05 $\pm$ 0.73
PromptHate	76.76 $\pm$ 0.95	67.82 $\pm$ 1.23	76.21 $\pm$ 1.05	68.08 $\pm$ 0.58	87.51 $\pm$ 0.74	79.38 $\pm$ 1.72
BLIP	76.80 $\pm$ 2.37	69.20 $\pm$ 1.84	80.59 $\pm$ 0.87	71.84 $\pm$ 1.11	87.09 $\pm$ 1.46	81.81 $\pm$ 1.74
ALBEF	79.40 $\pm$ 0.53	70.58 $\pm$ 0.50	83.24 $\pm$ 0.93	72.77 $\pm$ 1.00	85.49 $\pm$ 1.23	80.99 $\pm$ 0.80
Pro-CapBERT	77.50 $\pm$ 0.58	68.14 $\pm$ 0.64	79.62 $\pm$ 0.91	71.06 $\pm$ 0.88	89.04 $\pm$ 1.00	82.06 $\pm$ 1.92
Pro-CapPromptHate	<b>80.87</b> $\pm$ 0.66	<b>72.28</b> $\pm$ 0.90	82.53 $\pm$ 0.49	<b>73.06</b> $\pm$ 0.82	<b>90.25</b> $\pm$ 0.54	<b>83.25</b> $\pm$ 1.00

**Table 5: Model comparison with augmenting the image entities and demographic information.**

Dataset Model	FHM		MAMI		HarM	
	AUC.	Acc.	AUC.	Acc.	AUC.	Acc.
VisualBERT COCO	72.56 $\pm$ 0.80	64.28 $\pm$ 1.27	80.84 $\pm$ 0.67	72.86 $\pm$ 0.71	82.96 $\pm$ 0.98	78.81 $\pm$ 0.80
ViLBERT CC	75.72 $\pm$ 0.91	68.24 $\pm$ 0.44	80.33 $\pm$ 1.01	71.75 $\pm$ 1.14	84.79 $\pm$ 1.23	81.39 $\pm$ 1.62
MOMENTA	69.17 $\pm$ 4.71	61.34 $\pm$ 4.89	81.68 $\pm$ 2.80	72.10 $\pm$ 2.90	86.32 $\pm$ 3.83	80.48 $\pm$ 1.95
DisMultiHate	79.89 $\pm$ 1.71	71.26 $\pm$ 1.66	80.08 $\pm$ 0.55	71.87 $\pm$ 0.47	86.39 $\pm$ 1.17	81.24 $\pm$ 1.04
PromptHate	81.45 $\pm$ 0.74	72.98 $\pm$ 1.09	79.95 $\pm$ 0.66	70.31 $\pm$ 0.64	90.96 $\pm$ 0.62	84.47 $\pm$ 1.75
BLIP	76.40 $\pm$ 1.49	69.29 $\pm$ 1.44	80.63 $\pm$ 1.05	70.62 $\pm$ 1.48	86.88 $\pm$ 1.15	82.66 $\pm$ 1.13
ALBEF	80.77 $\pm$ 0.81	71.70 $\pm$ 0.98	82.45 $\pm$ 0.85	72.45 $\pm$ 0.96	86.91 $\pm$ 0.72	81.78 $\pm$ 1.20
Pro-CapBERT	79.75 $\pm$ 1.15	71.28 $\pm$ 0.91	81.20 $\pm$ 0.69	71.80 $\pm$ 1.42	89.75 $\pm$ 1.49	82.71 $\pm$ 1.60
Pro-CapPromptHate	<b>83.58</b> $\pm$ 0.60	<b>75.10</b> $\pm$ 0.97	<b>83.77</b> $\pm$ 0.75	<b>73.63</b> $\pm$ 0.75	<b>91.03</b> $\pm$ 1.51	<b>85.03</b> $\pm$ 1.51

**Table 6: Ablation study about the impact from the length of VQA answers.**

Ans. Length	FHM	MAMI	HarM
No Centric	70.08 $\pm$ 1.57	72.78 $\pm$ 0.63	80.11 $\pm$ 1.14
Penalty = 1	71.94 $\pm$ 0.97	73.06 $\pm$ 0.82	82.09 $\pm$ 1.21
Penalty = 2	72.28 $\pm$ 0.90	72.91 $\pm$ 1.16	82.85 $\pm$ 1.51
Penalty = 3	71.40 $\pm$ 1.06	72.47 $\pm$ 0.74	83.25 $\pm$ 1.00
Pro-CapPromptHate	<b>72.28</b> $\pm$ 0.90	<b>73.06</b> $\pm$ 0.82	<b>83.25</b> $\pm$ 1.00

While Pro-CapBERT does not out-perform ALBEF and BLIP all the time, performance is reasonably good given that in terms of trainable parameters, Pro-CapBERT is three times smaller than these two pre-trained models. Meanwhile, Pro-CapBERT shows even better results than the two models on HarM. Notably, HarM is a real-world dataset which is much noisier than FHM. HarM also focuses on a relatively new topic (COVID-19), which may not have been observed a lot by the two pre-trained models.

When comparing BLIP and ALBEF with PromptHate, which has a similar model size, PromptHate with Pro-Cap demonstrates significant advantages over the two models on three benchmarks, especially on the noisy HarM dataset. We conjecture that a possible reason is that multimodal pre-trained models leverage pre-training

data that is relatively cleaner, on a smaller scale and primarily comprises of non-memes. This leads to some difficulties when confronted with noisy real-world memes. In contrast pure language models are pre-trained on larger and noisier data, which may lead to some intrinsic robustness. If visual signals are reasonably converted to text, pure textual models can be competitive for multimodal tasks such as hateful meme detection.

Reinforcing the point of proper visual signal conversion, the enhanced performance of Pro-CapPromptHate over PromptHate highlights the importance of our probing-based captioning method, which provides essential cues for hateful content detection. With probe-based captioning, Pro-CapPromptHate is able to conduct deep multimodal reasoning that require background knowledge (due to the good performance on FHM), is stable towards noisy real-world meme data (according to performance on HarM), and has great generalization in meme detection (according to the good performance on all three benchmarks).

**With augmented image tags:** For a fair comparison with recent state-of-the-art models, we consider testing our proposed probe-captioning method with the same set of augmented image tags from baselines. To utilize the augmented image tags, we simply pad these tags at the end of each textual meme representation in a similar manner to [2]. With additional image information such



**Table 7: Comparison between Pro-Cap<sub>PromptHate</sub> with basic PromptHate. The image caption used by basic PromptHate is denoted as basic caption. Incorrect prediction in red. The content in (·) of the ground-truth is the target of the hateful meme.**

<b>Meme</b>			
<b>Ground Truth</b>	Hateful (religion)	Hateful (religion)	Hateful (race)
<b>Basic PromptHate</b>	Non-hateful	Non-hateful	Non-hateful
<b>Pro-Cap<sub>PromptHate</sub></b>	Hateful	Hateful	Hateful
<b>Meme text</b>	changing every single country it touches	no that is not his daughter that is his wife yet the world is silent	the definition of utter disgust in plain black and white
<b>Basic caption</b>	mughal structure is one of the largest mosques in the world.	portrait of a father hugging his daughter while smiling at camera in the living room at home.	love is in the air!.
<b>Pro-Cap</b>	(Content:·) a black cat sitting on a blue and white tiled floor. (Race:·) a black person is standing on a blue and white tiled floor in islamic. (Gender:·) a man in a black shirt is standing on a blue and white tiled floor with a clock on top of his head. (Country:·) islamic. (Religion:·) the person is a muslim and he is wearing a black t-shirt and a black sleeveless.	(Content:·) a man and a woman hugging on a couch. (Race:·) a white man and a white woman hugging on a white couch. (Gender:·) a man and a woman hugging on a white couch. (Country:·) islamic. (Religion:·) an muslim man and woman hugging on a white couch.	(Content:·) a black and white photo of a man and a woman. (Race:·) a black man and a white woman hugging in a black and white photo. (Gender:·) a man and a woman in a black and white photo. (Country:·) afghanistan. (Religion:·) he is a christian.



as entities and demographic information, most models have some improvements. An interesting thing is that neither BLIP nor ALBEF benefits much from additional image tags. This is because the additional tags are usually single words or short phrases, which may be noisy or redundant, while BLIP and ALBEF may be less capable of dealing with noisy inputs. Similar to the results in Table 4, when augmenting image information: 1) the simple Pro-Cap<sub>BERT</sub> still obviously surpasses multimodal pre-trained BERT-base models such as VisualBERT or ViLBERT; 2) the Pro-Cap<sub>BERT</sub> performs better than models with similar sizes but specifically designed for hateful meme detection (i.e., MOMENTA or DisMultiHate) in most cases; 3) the Pro-Cap<sub>BERT</sub> achieves comparable results compared with more powerful multimodal pre-trained models, which is about three times larger and surpasses them on the HarM dataset, which is real-world and noisy; 4) Pro-Cap<sub>PromptHate</sub> surpasses the original PromptHate and achieves the best performance on three benchmarks as well. An interesting point is that comparing Pro-Cap<sub>PromptHate</sub> without any augmented tags and original PromptHate with augmented additional image information, they achieve comparable performance on FHM and HarM and the former even surpasses the latter on MAMI. However, extracting the additional image information is expensive and laborious, which can be replaced by probing-based captioning according to the experimental results. The equally good performance on three benchmarks highlights the stability and generalization of our proposed approach.

## 5.4 Ablation Study

In this section, we conduct ablation studies to better understand our Pro-Cap method. Specifically, we consider the impact of asking different questions and the impact of the length of answers to the probing questions. To eliminate other factors, we consider Pro-Cap<sub>PromptHate</sub> without any augmented image tags. For brevity, we only show accuracy in this section. We present the full results in Appendix.

**The impact of asking hateful-content centric questions:** We first conduct an ablation study on the effect of prompting PVLMS with questions facilitating hateful meme detection. According to Table 2, the first question asks about the image content while all questions in the second block are for common vulnerable targets of hateful contents. To better understand the impact of including image captions generated by these target-specific questions, we experiment with a setting where captions from the target-specific questions are removed and only the generic caption about image content is used. The results are shown in the first block of Table 6. Compared with the last block of the table, we observe that with captions generated by target-specific probing questions, the model’s performance improved on all three datasets, specifically with over 2% on FHM and over 3% on HarM. However, we notice minor improvement on MAMI. We believe that this is because MAMI memes are all related to woman and generic captions about meme images may already cover the gender of persons in the image. However, the other two datasets involve memes with more complexities and

**Table 8: Error cases of Pro-CappromptHate.**

<b>Meme</b>		
<b>GT</b>	Hateful (gender)	Non-hateful
<b>Pred</b>	<b>Non-hateful</b>	<b>Hateful</b>
<b>Meme text</b>	scientist are working hard to cure them all	islam is a religion of peace stop criticizing my religion
<b>Pro-Cap</b>	(Content:) two women in wedding dresses kissing each other. (Race:) a white woman kissing a brunette woman in a wedding dress. (Gender:) a woman is kissing a man in a wedding dress. (Country:) the person in the image comes from a country in the philippines. (Religion:) the person in the image is a christian.	(Content:) a man with a beard laughing in the woods. (Race:) a african man with a beard and a red hat is smiling in the woods. (Gender:) a man with a beard and a red hat in front of a wooded area. (Country:) egypt is the country that the person in the image comes from. (Religion:) he is a muslim man with a beard and a red tiara on his head.

therefore asking a wide ragen of target-specific probing questions is more helpful. It also implies that in real-world hateful meme detection, probing-based captioning would be helpful.

**The length of answers to probing questions:** We apply BLIP-2 as a zero-shot VQA model. Different from existing VQA benchmarks [7, 10], where answers are often single words or short phrases, we may want the answers used as image captions to be longer and thus more informative. In this cases, we experiment with answers of different length. To conduct the analysis, we set the length penalty in BLIP-2’s text decoder for answer generation with different values (i.e., 1, 2 and 3). With increased length penalty, longer answers are encouraged. We show results of model performance with different answer length in Table 6. The results show that detection performance is robust and does not vary much with different answer lengths. This indicates the stability of the Pro-Cap method. On the other hand, to a very small extent, different datasets do favor answers of different lengths. For instance, the HarM dataset prefers longer answers while the MAMI dataset prefers shorter answers.

## 5.5 Case Study

In this section, we conduct case studies to better understand the strengths and limitations of our proposed method. We first compare Pro-CappromptHate against PromptHate with image captions and show examples in Table 7. From the three examples, we observe that in most cases, generic captions about the image content do not provide the key information for hateful meme detection, while asking questions about common vulnerable targets helps. For instance,

in the first example, the answer from asking questions about race, country and religion all provide some key words such as *islamic* or *muslim*; in the second example, answers to questions about country and religion are important image captions and the answer to the race-related question is the most important for hateful meme detection. In contrast, we observe that the basic captions in the original PromptHate miss these crucial facts about the meme images.

Next, we conduct error analysis about our proposed probe-captioning in Table 8. In the first example, all probe-captions generate sufficient image captions for the hateful meme detection, while the model still fails at prediction. This may be due to the current language models performing poorly in further complex reasoning. We also note that the small scale of hateful meme datasets may be inadequate for training a model to perform complex reasoning. Recent studies about large language models pre-trained with trillions of words [33] may facilitate hateful meme detection to some extent. Besides, we observe minor errors in predicted answers from the zero-shot VQA model (e.g., the wrong prediction of “a woman kissing a man” when asking about gender). It highlights that with the development of better zero-shot VQA models, the our strategy could potentially facilitate more for the two text-based hateful meme detection models. The second example highlights a limitation of most hateful content detection models in that they may be biased. During the training stage, there may be hateful contents towards Muslims so that once models seen Muslims, they tend to predict the meme as hateful. To alleviate the issue, debiasing techniques may be needed. Due to space limitation, we omit visualization examples in the main pages and refer the reader to examples in Appendix.

## 6 CONCLUSION

In this study, we attempt to leverage pre-trained vision-language models (PVLMs) in a low-computation-cost manner to aid the task of hateful meme detection. Specifically, without any fine-tuning of PVLMs, we probe them in a zero-shot VQA manner to generate hateful content-centric image captions. With the distilled knowledge from large PVLMs, we observe that a simple language model, BERT, can surpass all multimodal pre-trained BERT models of a similar scale. PromptHate with probe-captioning outperforms previous results significantly and achieves the new state-of-the-art on three benchmarks.

**Limitations:** We would like to point out a few limitations of the proposed method, suggesting potential future directions. Firstly, we heuristically use answers to all probing questions as Pro-Cap, even though some questions may be irrelevant to the meme target. We report the performance of PromptHate with the answer from one probing question in Appendix, highlighting that using all questions may not be the optimal solution. A future direction could involve training a model to dynamically select probing questions that are most relevant for meme detection. Secondly, although we demonstrate the effectiveness of Pro-Cap through performance and a case study in this paper, more thorough analysis is needed. For instance, in the future, we could use a gradient-based interpretation approach [31] to examine how different probing questions influence the final results, thereby enhancing the interpretation of the models.

## REFERENCES

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *CoRR* (2022). arXiv:2204.14198
- [2] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 321–332.
- [3] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics EACL*. 2136–2148.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4171–4186.
- [5] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL*. 533–549.
- [6] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, ACL/IJCNLP*. 3816–3830.
- [7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 6325–6334.
- [8] Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the Underlying Meaning of Multimodal Hateful Memes. *arXiv preprint arXiv:2305.17678* (2023).
- [9] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*. 3651–3655.
- [10] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 6700–6709.
- [11] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS Workshop*.
- [12] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems, NeurIPS*.
- [13] Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. *CoRR* abs/2210.05916 (2022).
- [14] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling Hate in Online Memes. In *MM '21: ACM Multimedia Conference*. 5138–5147.
- [15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR* (2023). arXiv:2301.12597
- [16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML*, Vol. 162. 12888–12900.
- [17] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *CoRR* (2021). arXiv:2107.07651
- [18] 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. *CoRR* (2019). arXiv:1908.03557
- [19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV European Conference*, Vol. 8693. 740–755.
- [20] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A Multimodal Framework for the Detection of Hateful Memes. *arXiv preprint arXiv:2012.12871* (2020).
- [21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* (2019). arXiv:1907.11692
- [22] Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR* (2017). arXiv:1711.05101
- [23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR* (2019). arXiv:1908.02265
- [24] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOA5 Shared Task on Fine Grained Hateful Memes Detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 201–206.
- [25] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *CoRR* (2021). arXiv:2111.09734
- [26] Niklas Muennighoff. 2020. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *CoRR* (2020). arXiv:2012.07788
- [27] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tammy Chakraborty. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*. 2783–2796.
- [28] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tammy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP*. 4439–4455.
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning, ICML*, Vol. 139. 8748–8763.
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.
- [31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision, ICCV*. 618–626.
- [32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL*. 2556–2565.
- [33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* (2023). arXiv:2302.13971
- [34] Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *CoRR* (2020). arXiv:2012.12975
- [35] Yi Zhou and Zhenhao Chen. 2020. Multimodal Learning for Hateful Memes Detection. *arXiv preprint arXiv:2011.12870* (2020).
- [36] Jiawen Zhu, Roy Ka-Wei Lee, and Wen Haw Chong. 2022. Multimodal zero-shot hateful meme detection. In *Proceedings of the 14th ACM Web Science Conference 2022*. 382–389.
- [37] Ron Zhu. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *CoRR* (2020). arXiv:2012.08290