# Learning Multimodal Contrast with Cross-modal Memory and Reinforced Contrast Recognition

**Yuanhe Tian**[♠♥], **Fei Xia**[♥], **Yan Song**[♠†]
[♠]University of Science and Technology of China [♥]University of Washington
[♥]{yhtian, fxia}@uw.edu [♠]clksong@gmail.com

## Abstract

In many practical scenarios, contents from different modalities are not semantically aligned; for instance, visual and textual information may conflict with each other, resulting in non-compositional expression effects such as irony or humor. Effective modeling and smooth integration of multimodal information are crucial for achieving good understanding of the contrast across modalities. Being focusing on image-text matching, most current studies face challenges in identifying such contrast, leading to limitations in exploring the extended semantics when images and texts do not match. In this paper, we propose an LLM-based approach for learning multimodal contrast following the encoding-decoding paradigm, enhanced by a memory module with reinforced contrast recognition, and use a series of tasks that have the nature of multimodal contrast to verify our approach. The memory module learns the integration between visual and textual features with trainable memory vectors and the reinforced contrast recognition uses self-rejection sampling to optimize the memory to further enhance learning multimodal contrast. The resulted information, accompanied with visual and text features, is finally fed into the LLM to predict corresponding labels. We experiment our approach on four English and Chinese benchmark datasets, where it outperforms strong baselines and state-of-the-art studies.[1]

## 1 Introduction

Multimodal information have become a widespread form of expression in many real-world applications, such as news feeding, social media, and instance messaging, etc., (Kiela et al., 2020; Gomez et al., 2020; Sharma et al., 2020; Suryawanshi et al., 2020; Li et al., 2022), where in most cases they are fused with image and text pairs. In order to enhance the expressive effect, images and texts are not always semantically aligned, resulting in irony or humor expressions, causing hateful or joyful emotions spreading around groups of people. Such misalignment has the non-compositional effect that similar to idioms, where the overall meaning is not the combination of the meanings from its components.[2] For example, Figure 1 shows three multimodal memes: (a) is a hateful meme; (b) and (c) are not. Note that meme (a) has the same image as (c) and the same text as (b). In other words, a meme can be hateful even when neither its image nor its text is. This example demonstrates the importance of understanding the relationship between image and text, as the hateful attitude in a meme arises from the multimodal contrast[3], which allows the integrated information from different modalities to convey a message (e.g., hateful information) that cannot be expressed by any single modality alone.

Existing approaches for multimodal understanding mainly focus on image-text matching (Vinyals et al., 2016; Li et al., 2019; Chen et al., 2020; Qin and Song, 2022; Park and Paik, 2023; Ramos et al., 2023; Wang et al., 2023a) and utilize advanced visual and text encoders (such as CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), FLAVA (Singh et al., 2022), and SLIP (Mu et al., 2022), etc.) to extract multimodal features, and subsequently align or fuse them by vector concatenation, outer production, or attentions to perform downstream tasks, such as hateful meme detection (HMD) (Kiela et al., 2019; Li et al., 2019; Radford et al., 2021; Goyal et al., 2022; Kumar and Nandakumar, 2022; Koutlis et al., 2023). These HMD models are good at identifying hateful

[†]Corresponding author.

[1]Materials related to the paper is available at https://github.com/synlp/MemRCRHMD.

[2]For example, the idiom "*crossing the Rubicon*" means "passing a point of no return", where every word in this idiom does not present such meaning.

[3]The term *contrast* in this work represents the aforementioned non-compositional effect when combining information from multiple modalities.



*Findings of the Association for Computational Linguistics: ACL 2024*, pages 6561–6573


(a) Hateful (b) Non-hateful (c) Non-hateful

Figure 1: Three memes: (a) is hateful; (b) and (c) are not. Here, (a) and (b) share the same text, and (a) and (c) share the same image. This example shows that a meme can be hateful even when neither of its image nor text is.

memes when images or text present explicit biases, but are unable to effectively recognize hateful information that is derived from the contrast between image and text. Although there are efforts in utilizing additional resources or using model ensemble to improve tasks like HMD (Muennighoff, 2020; Lippe et al., 2020; Velioglu and Rose, 2020; Zhu, 2020; Cao et al., 2023b), they mainly enhance the generalization ability through more training data or ensemble of multiple models, overlooking the contrast between multiple modalities that lead to better understanding of the non-compositional effect of cross-modal information fusion.

In this paper, we propose an approach with LLM to learn multimodal contrast through cross-modal memory and reinforced contrast recognition (RCR). The cross-modal memorizing module learns how to capture the information from multiple modalities, and the reinforced contrast recognition utilizes self-rejection training to enhance the memory in learning the contrast by further optimizing the loss function. We use a series of representative tasks that require to understand the contrast from multimodalities to verify our approach.[4] Evaluations on four benchmark datasets show that our approach outperforms strong baselines and existing approaches, demonstrating the benefits of memory and reinforcement learning with self-rejection training for all the tasks.

## 2 The Approach

Figure 2 illustrates the framework of our approach, which follows the encoding-decoding paradigm to perform multimodal classification, which predicts a label $\widehat{\mathcal{Y}}$ based on the image $\mathcal{V}$ and embedded text $\mathcal{T}$ in a given pair $(\mathcal{V}, \mathcal{T})$. It contains three essential components: the backbone model, the cross-modal memory, and the RCR. The memory module is inserted in between the visual encoding and LLM decoding in the backbone, illustrated at the top of Figure 2. The RCR, demonstrated at the bottom of Figure 2, enhances the memory by enhancing the memory module with contrast information via self-rejection training, where an additional term is added to the loss function. In the following subsections, we firstly illustrate the backbone model, then the cross-modal memory, and finally presents the RCR.

### 2.1 The Backbone Model

The encoding and decoding processes are two essential components in the backbone model. Specifically, the visual encoding process ($f_{ve}$) extracts salient features from the input image and LLM decoding ($f_d$) utilizes the multimodal information to predict the final classification label $\widehat{\mathcal{Y}}$.

**Visual Encoding** Our visual encoder, following the procedure of BLIP2 (Li et al., 2023), has three components: the vision Transformer $f_v$ (Dosovitskiy et al., 2021), the Q-Former $f_q$ (Li et al., 2023), and a linear projection layer. The three modules are sequentially interconnected to extract visual feature $\mathbf{v}$ from the input meme $\mathcal{V}$ through

$$\mathbf{v} = f_{ve}(\mathcal{V}) = Linear(f_q(f_v(\mathcal{V}))) \quad (1)$$

The vision Transformer $f_v$ extracts crucial visual features from the meme. The Q-Former $f_q$ translates these features into a textual semantic space. Finally, the linear projection layer transforms the resulted representation into latent vector $\mathbf{v}$ that is used in the subsequent processes.

**LLM Decoding** Existing studies on LLM have demonstrated the significant impact of prompting

[4] Our approach is able to be applied to other similar tasks depending on modeling the contrast among multimodalities.

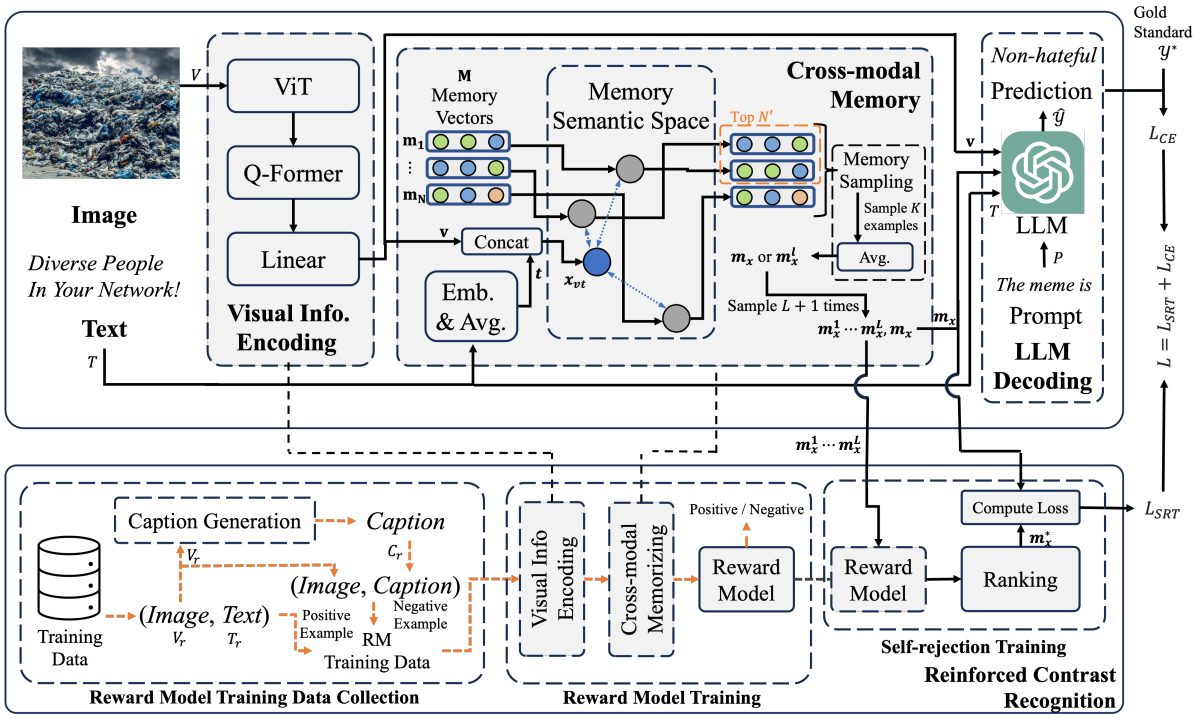



Figure 2: The overall architecture of our approach using HMD as a demonstration example. The top part illustrates our backbone model with cross-modal memory, and bottom part shows the RCR process. The workflow of collecting data and training the reward model is illustrated in orange arrows in the RCR part. Then the trained reward model is used in the self-rejection training to optimize the model, illustrated at the right bottom part. Visual encoding and cross-modal memory are shared by the backbone model in the main process and the reward model training process, which is marked by the boxes with dashed black lines.

on model performance (Brown et al., 2020; Lester et al., 2021; Ouyang et al., 2022; Liu et al., 2022). For better prompting, we use the visual feature $\mathbf{v}$ and the contrast vector $\mathbf{m}_x$ obtained from cross-modal memorizing as soft prompts to instruct our LLM for final classification. Specifically, we feed $\mathbf{v}$, $\mathbf{m}_x$, as well as the original text $\mathcal{T}$, into the LLM to determine the label $\widehat{\mathcal{Y}}$, e.g., *hateful* or *non-hateful* if the task is HMD. A prompt $P$ is required to instruct the LLM to process the input and predict the label. For example, we design a simple prompt, i.e., "*The meme is* __" for HMD, which instructs LLM to fill in the blank with "*hateful*" or "*non-hateful*".[5] We feed $\mathbf{v}, \mathbf{m}_x, \mathcal{T}, P$ into our LLM (e.g., Vicuna (Chiang et al., 2023)) and obtain the hidden vector $\mathbf{h}$ from its last layer by

$$\mathbf{h} = LLM(\mathbf{v}, \mathbf{m}_x, \mathcal{T}, P) \quad (2)$$

Afterwards, we use a projection layer to map $\mathbf{h}$ into the output label space and use a *softmax* classifier to predict the class label $\widehat{\mathcal{Y}}$ of the input image-text pair following the standard classification process. In training, we compare the prediction $\widehat{\mathcal{Y}}$ with the gold standard $\mathcal{Y}^*$ and compute the cross-entropy loss $\mathcal{L}_{CE}$ to optimize the model.

## 2.2 Cross-modal Memory

Memory mechanism is demonstrated to be effective in modeling task-related information (Song et al., 2018; Nie et al., 2020; Tian et al., 2022, 2023a) The memory module is designed to better fusing mutlimodal features, where we use a memory matrix $\mathbf{M}$, stacked with $N$ memory vectors (denoted by $\mathbf{M} = [\mathbf{m}_1, \cdots, \mathbf{m}_N]$) to record such feature fusion. Each memory vector is interpreted as a potential aspect in multimodal fusion that results in a particular type of information, e.g., hateful information for the HMD task or contrast that causes humor. This module has three main components: text encoding, memory ranking, and sampling.

Text encoding obtains text representations to facilitate subsequent processes. We follow the standard approach to use a word embedding matrix to map all words in the text $\mathcal{T}$ into their corresponding word embeddings. Then, we compute the average of the embeddings of all words in the input text and denote the resulted vector $\mathbf{t}$ as the text features.

Memory ranking locates relevant memory vec-

[5]The prompt may change for different tasks.

tors according to the encoded multimodal information and assigns appropriate scores to them. We concatenate visual and text features and obtain the multimodal feature $\mathbf{x}_{vt} = \mathbf{v} \oplus \mathbf{t}$. Then, we compute the score $s_n$ that measures the semantic similarity between the $n$-th memory vector $\mathbf{m}_n$ and $\mathbf{x}_{vt}$ by

$$s_n = \frac{\exp(\mathbf{x}_{vt} \cdot \mathbf{W}^M \cdot \mathbf{m}_n)}{\sum_{n=1}^{N} \exp(\mathbf{x}_{vt} \cdot \mathbf{W}^M \cdot \mathbf{m}_n)} \quad (3)$$

where $\mathbf{W}^M$ is a trainable parameter matrix to align $\mathbf{m}_n$ and $\mathbf{x}_{vt}$. Finally, we rank all memory vectors in descending order based on their scores and select the top $N'$ vectors (denoted as $\mathbf{m}_{n_1} \cdots \mathbf{m}_{n_{N'}}$) as the relevant vectors for later processing.

Memory sampling further processes memory vectors and outputs a vector $\mathbf{m}_x$ that carries the essential multimodal fusion information between visual and text features for later steps. In detail, we normalize the scores of the relevant vectors and randomly select $K$ vectors from $\mathbf{m}_{n_1} \cdots \mathbf{m}_{n_{N'}}$ (repetition of the same vector is allowed) based on their scores, where higher scores lead to better chance to be selected. We then average the select vectors and obtain the output contrast vector $\mathbf{m}_x$ by

$$\mathbf{m}_x = \frac{1}{K} \sum_{k=1}^{K} s_{n_k} \mathbf{m}_{n_k} \quad (4)$$

where $s_{n_k}$ is the score for $\mathbf{m}_{n_k}$ obtained by Eq. (3) and $\mathbf{m}_x$ is used as input in LLM decoding.

### 2.3 Reinforced Contrast Recognition

The goal of RCR is to help the cross-modal memory module in producing a better $\mathbf{m}_x$. However, we do not have a gold standard for $\mathbf{m}_x$. Therefore, we need to create a silver standard $\mathbf{m}_x^*$ and add the difference between $\mathbf{m}_x$ and $\mathbf{m}_x^*$ to the loss function. To create $\mathbf{m}_x^*$, we propose to use a reward model to select $\mathbf{m}_x^*$ from a candidate list $\mathbf{m}_x^1, \cdots, \mathbf{m}_x^L$. In order to generate the candidate list, we repeat the sampling and averaging process illustrated in Eq. (4) for $L$ times and obtain a list of different vectors $\mathbf{m}_x^1 \cdots \mathbf{m}_x^L$. Specifically, there are three main steps in RCR: reward model training data collection, reward model training, and self-rejection training. In the first and second steps, we collect data and train a reward model to rank $\mathbf{m}_x^1 \cdots \mathbf{m}_x^L$ based on their effectiveness in representing the contract information between multiple modalities. Then we perform self-rejection training by using the most effective contrast vector $\mathbf{m}_x^*$ to optimize $\mathbf{m}_x$ so that $\mathbf{m}_x$ is trained to be closer to $\mathbf{m}_x^*$. Details are presented as follows.

**Reward Model Training Data Collection** The goal of the reward model is to assess whether the encoded vectors from the memory module contain contrast information, and we collect positive and negative examples to train it. Therefore, we rely on the training examples for a running task that is based on such contrast (e.g., HMD) to serve as positive examples; then take the ordinary (image, caption) pairs from image captioning tasks as negative examples since the images and their corresponding captions generally share similar semantics. As a result, we randomly select instances, i.e., image-text pairs $(\mathcal{V}_r, \mathcal{T}_r)$, from the training data of particular tasks as a positive examples. Then we generate captions $\mathcal{C}_r$ for images $\mathcal{V}_r$ using an off-the-shelf image captioning toolkit and combine with their image to form negative examples $(\mathcal{V}_r, \mathcal{C}_r)$.

**Reward Model Training** In training the reward model, we apply the same visual encoding and the memory module in our approach to compute the contrast vectors for the positive and negative samples by $\mathbf{v}_m^{pos} = f_m(f_{ve}(\mathcal{V}_r), \mathcal{T})$ and $\mathbf{v}_m^{neg} = f_m(f_{ve}(\mathcal{V}_r), \mathcal{C})$, where $\mathbf{v}_m^{pos}$ and $\mathbf{v}_m^{neg}$ denote the positive and negative contrast vectors, respectively, and $f_m$ means the memory module. Finally, we feed $\mathbf{v}_m^{pos}$ and $\mathbf{v}_m^{neg}$ to the reward model $f_r$, which is a multi-layer perceptron, and compute the reward (denoted as $r_{pos}$ and $r_{neg}$, respectively) for the vectors by $r_{pos} = sigmoid(f_r(\mathbf{v}_m^{pos}))$ and $r_{neg} = sigmoid(f_r(\mathbf{v}_m^{neg}))$, and compute the loss $\mathcal{L}_r$ to optimize the reward model by

$$\mathcal{L}_r = -log(r_{pos}) - log(1 - r_{neg}) \quad (5)$$

**Self-rejection Sampling** In this step, we use the reward model to reject ineffective vectors and choose the best one $\mathbf{m}_x^*$ to improve $\mathbf{m}_x$, which is similar to the process used in Touvron et al. (2023b). In doing so, we feed all memory sampled vectors $\mathbf{m}_x^1 \cdots \mathbf{m}_x^L$ to the reward model $f_r$ and compute the reward for each of them, and select the vector $\mathbf{m}_x^*$ with the highest reward score and use it as the gold standard to assess whether a sampled vector from the memory module is good enough to carry essential task-specific contrast information for final classification. Finally, we compute the loss

$$\mathcal{L}_{SRT} = |\mathbf{m}_x^* - \mathbf{m}_x| \quad (6)$$

and add it to $\mathcal{L}_{CE}$ to get the final loss $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{SRT}$ to update the entire framework.

| | HMC | | | Memotion7k | | MultiOFF | | | Memeplate | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Dev | Test | Train | Test | Train | Dev | Test | Train | Dev | Test |
| # of Meme | 8,500 | 500 | 1,000 | 6,992 | 1,879 | 445 | 149 | 149 | 3,746 | 700 | 738 |
| Avg. Tokens Per Meme | 11.7 | 10.2 | 10.4 | 14.7 | 15.7 | 41.4 | 45.2 | 47.0 | 20.3 | 20.4 | 20.0 |

Table 1: Statistics of experiment datasets, where the number of meme and the average number of tokens (i.e., words for English and characters for Chinese) for each meme are reported.

## 3 Experiment Settings

### 3.1 Datasets

For our experiments, we employ three English datasets, namely, hateful meme challenge (HMC) dataset (Kiela et al., 2020), Memotion7K (Sharma et al., 2020), and MultiOFF (Suryawanshi et al., 2020), and one Chinese dataset named Memeplate (Li et al., 2022). These datasets cover a wide range of tasks that require modeling of contrasts. Specifically, HMC is for HMD. Memotion7k contains three tasks: sentiment classification (T1), humor classification (T2), and scales of semantic classes (T3). MultiOFF is designed for detecting offensive content from image-text pairs. Memeplate is for multimodal humor recognition. We use the official training, development, and test data split of all datasets. Herein, for HMC, we follow the convention of most existing studies (Radford et al., 2021; Singh et al., 2022; Cao et al., 2023b; Koutlis et al., 2023) to evaluate model performance on the development set. The statistics of the datasets are reported in Table 1, where the number of memes and the average number of tokens (i.e., words for English and characters for Chinese) for each meme are presented.

### 3.2 Baselines

We run baselines with small language models and LLMs as the backbones following the BLIP2 (Li et al., 2023) architecture. For small language models, we employ GPT-2 (Radford et al., 2019). For LLMs, we use MiniGPT-4 (which is demonstrated to be effective in many multimodal tasks) for English and use Ziya-BLIP2-Visual (Zhang et al., 2022a) for Chinese processing. Based on small and large models, our experiments include three baselines. The first is the vanilla BLIP2 with small and large language models. The second baseline (i.e., "+M") adds the proposed memory module on top of the first one. The third baseline (i.e., "+RCR") adds reinforced contrast recognition (RCR) on top of the first one. We concatenate visual and text features and use the resulting vector (i.e., $\mathbf{x}_{vt}$) to represent the contrast vector (i.e., $\mathbf{m}_x$) and randomly set 33% values in $\mathbf{m}_x$ to zero to facilitate RCR training.

### 3.3 Implementation Details

We use the default settings of BLIP2 (with GPT-2), MiniGPT-4, or Ziya-BLIP2-Visual, which contain visual encoding and LLM decoding processes. For the visual encoding process, we follow the standard architecture using visual transformer and Q-Former, which contain 40 and 12 layers of multi-head attentions, respectively. For the LLM decoding process, the LLMs in BLIP2 (with GPT-2), MiniGPT-4, and Ziya-BLIP2-Visual utilize 12, 32, and 40 layers of Transformers, respectively.

In training our approach, we alternate between the following two procedures for every 100 steps: (1) updating the parameters of different components in visual encoding, memory module, and LLM using the cross-entropy loss from comparing the predicted labels with gold standards and (2) updating the reward model and the memory module through RCR.[6] For evaluation, we follow existing studies (Kiela et al., 2020; Li et al., 2022; Cao et al., 2023b; Koutlis et al., 2023) to use accuracy and AUROC for HMC, accuracy and F1 for MultiOFF, F1 for Memotion7K, and accuracy and F1 for Memeplate. For the hyper-parameters, we set the numbers of memory vectors (i.e., $N$) to 200 for HMC and 150 for other datasets. For all datasets, we use 20 as the memory sampling size (i..e., $K$), and 4 as the sampling time $L$. We set learning rate to $1 \times 10^{-6}$ with a batch size of 32. For other hyper-parameters, we tune them on the development set[7] and select the ones with the best performance to train models and evaluate them on the test sets. We run all models five times using different random seeds and report the average and standard deviation of their performance.

[6] For the third baseline with RCR (i.e., +RCR), we update the parameters of visual encoding and the token embeddings of the input text during training, so as to appropriately work with the absence of the memory module.

[7] For HMC, we randomly select 10% of the training data and use it to tune hyper-parameters.

| | HMC | | MEMOTION7K | | | MULTIOFF | | MEMEPLATE | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | AUROC | T1 | T2 | T3 | ACC | F1 | ACC | F1 |
| **GPT-2** | $73.28_{\pm 0.20}$ | $83.01_{\pm 0.22}$ | $35.09_{\pm 0.24}$ | $47.58_{\pm 0.20}$ | $32.10_{\pm 0.19}$ | $68.32_{\pm 0.24}$ | $61.67_{\pm 0.23}$ | $53.37_{\pm 0.19}$ | $47.09_{\pm 0.23}$ |
| **+M** | $74.00_{\pm 0.25}$ | $83.81_{\pm 0.20}$ | $35.50_{\pm 0.20}$ | $48.11_{\pm 0.23}$ | $32.85_{\pm 0.26}$ | $68.94_{\pm 0.20}$ | $62.48_{\pm 0.21}$ | $54.37_{\pm 0.26}$ | $47.76_{\pm 0.22}$ |
| **+RCR** | $74.52_{\pm 0.22}$ | $84.56_{\pm 0.20}$ | $36.12_{\pm 0.19}$ | $48.64_{\pm 0.19}$ | $33.48_{\pm 0.23}$ | $69.53_{\pm 0.21}$ | $63.05_{\pm 0.21}$ | $55.10_{\pm 0.26}$ | $48.86_{\pm 0.24}$ |
| **+M+RCR** | $^{*}\mathbf{75.08}_{\pm 0.23}$ | $^{*}\mathbf{84.91}_{\pm 0.20}$ | $^{*}\mathbf{38.06}_{\pm 0.21}$ | $^{*}\mathbf{50.01}_{\pm 0.21}$ | $^{*}\mathbf{34.57}_{\pm 0.22}$ | $^{*}\mathbf{71.58}_{\pm 0.25}$ | $^{*}\mathbf{64.02}_{\pm 0.19}$ | $^{*}\mathbf{55.86}_{\pm 0.22}$ | $^{*}\mathbf{49.75}_{\pm 0.25}$ |
| **LLM** | $76.20_{\pm 0.26}$ | $84.44_{\pm 0.23}$ | $37.48_{\pm 0.21}$ | $49.75_{\pm 0.21}$ | $33.76_{\pm 0.23}$ | $71.51_{\pm 0.20}$ | $64.87_{\pm 0.23}$ | $54.38_{\pm 0.25}$ | $47.97_{\pm 0.22}$ |
| **+M** | $76.56_{\pm 0.22}$ | $84.84_{\pm 0.25}$ | $38.82_{\pm 0.20}$ | $50.80_{\pm 0.23}$ | $34.83_{\pm 0.25}$ | $72.11_{\pm 0.20}$ | $65.94_{\pm 0.23}$ | $55.12_{\pm 0.21}$ | $48.70_{\pm 0.28}$ |
| **+RCR** | $77.01_{\pm 0.22}$ | $85.40_{\pm 0.25}$ | $40.82_{\pm 0.23}$ | $51.40_{\pm 0.22}$ | $35.61_{\pm 0.25}$ | $73.18_{\pm 0.26}$ | $67.72_{\pm 0.22}$ | $55.86_{\pm 0.23}$ | $49.39_{\pm 0.20}$ |
| **+M+RCR** | $^{*}\mathbf{77.88}_{\pm 0.24}$ | $^{*}\mathbf{86.34}_{\pm 0.23}$ | $^{*}\mathbf{41.56}_{\pm 0.21}$ | $^{*}\mathbf{52.73}_{\pm 0.21}$ | $^{*}\mathbf{35.88}_{\pm 0.24}$ | $^{*}\mathbf{74.09}_{\pm 0.20}$ | $^{*}\mathbf{68.43}_{\pm 0.21}$ | $^{*}\mathbf{56.52}_{\pm 0.20}$ | $^{*}\mathbf{50.21}_{\pm 0.21}$ |

Table 2: The average and standard deviation of the performance from various models on benchmark datasets. "GPT-2" and "LLM" stand for BLIP2 baseline models use small and large language models, respectively. "+M" and "+RCR" refer to that the memory module and the RCR are used on top of the baselines, respectively. Results marked by $^{*}$ means that the improvements are statistically significant at $p \leq 0.05$ level over all baselines.

| | ACC | AUROC |
|---|---|---|
| Muennighoff (2020) | - | 81.56 |
| Velioglu and Rose (2020) | 70.93 | 75.21 |
| Lippe et al. (2020) | - | 77.39 |
| Radford et al. (2021) | - | 77.30 |
| Goyal et al. (2022) | - | 73.40 |
| Kumar and Nandakumar (2022) | - | 81.55 |
| Singh et al. (2022) | - | 76.70 |
| Cao et al. (2023a) | 72.28 | 80.87 |
| Koutlis et al. (2023) | 73.60 | 80.10 |
| Cao et al. (2023b) | 72.98 | 82.45 |
| †△Liu et al. (2023) | 76.20 | 84.57 |
| **Ours** | **77.88** | **86.34** |

Table 3: Comparison of the average performance of our approach with the existing studies on the development set of HMC. "†" means the results are our own runs using their multimodal approaches. "△" indicates that LLMs are used to predict labels. The markups are the same for following tables.

| | T1-F1 | T2-F1 | T3-F1 |
|---|---|---|---|
| Keswani et al. (2020) | 35.5 | - | - |
| Vlad et al. (2020) | 34.5 | 51.8 | 31.7 |
| Guo et al. (2020) | 35.2 | 51.5 | 32.3 |
| Kumari et al. (2021) | 36.8 | - | - |
| †Ouaari et al. (2022) | 35.3 | - | - |
| Zhang et al. (2022b) | 36.6 | 46.9 | - |
| Zhong et al. (2022) | 37.0 | - | - |
| Koutlis et al. (2023) | 39.6 | 51.9 | 34.3 |
| **Ours** | **41.56** | **52.73** | **35.88** |

Table 4: Performance comparison of different models on the test set of three tasks on Memotion7k dataset.

| | ACC | F1 |
|---|---|---|
| Lee et al. (2021) | - | 64.6 |
| Zhong et al. (2022) | - | 67.1 |
| Koutlis et al. (2023) | 68.5 | 62.5 |
| **Ours** | **74.09** | **68.43** |

Table 5: Comparison of different models on the test set of MultiOFF dataset.

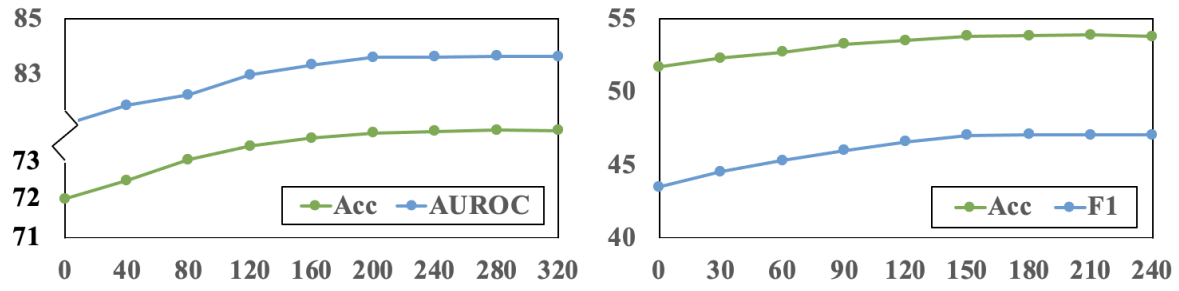
## 4 Results and Analysis

### 4.1 Overall Performance

The average performance with standard deviations of baselines and our approach for all datasets under different settings are reported in Table 2, with following observations. First, overall, our approach (i.e., +M+RCR) outperforms the vanilla BLIP2 (GPT-2), MiniGPT-4, and Ziya-BLIP2-Visual baselines, which indicates the effectiveness of our approach to learning contrast information for different tasks. Second, when the memory (i.e., "+M") or the RCR module (i.e., "+RCR") is added to the vanilla baseline, improvements are all observed, which is the evidence for the effectiveness of each individual module in capturing contrast between visual and textual data, thereby enhancing model performance.[8] Third, when comparing "+M" and "+RCR", we find that RCR consistently exhibits superior performance across various configurations, underscoring the advantage of discriminatively learning the contrast information. Fourth, our full model that integrates both the memory and RCR outperforms all baseline models, demonstrating the effectiveness of complementing each other.

We further compare our approach with existing studies for HMC, Memotion7K, MultiOFF, and Memeplate in Table 3-6, where the results demonstrate state-of-the-art performance. Particu-

[8]With the training data from particular tasks that contains such contrast, the memory module is also able to learn that information as that performed in the RCR process.




(a) HMC (b) Memeplate

Figure 3: Curves of model performance on HMC and Memeplate with respect to different numbers of memory vectors used in the memory module.

| | ACC | F1 |
|---|---|---|
| †Yang et al. (2022) | 52.57 | 46.21 |
| †△Yang et al. (2023) | 55.43 | 48.80 |
| †△Hu et al. (2023) | 55.08 | 48.97 |
| †△University (2023) | 55.76 | 49.49 |
| **Ours** | **56.52** | **50.21** |

Table 6: Performance comparison of different models on the test set of Memeplate dataset.

larly, our approach outperforms the ones that use advanced pre-trained models for image and text processing (Kumar and Nandakumar, 2022; Singh et al., 2022; Koutlis et al., 2023). The reason is that, these multimodal models generally perform the HMD, multimodal sentiment analysis, offensive content detection, and humor recognition in the same way as image captioning, therefore not focus on the contrast across modalities whereas image captioning emphasizes the content shared by these modalities. Compared with these studies, the performance of our approach on all tasks confirm the validity of explicitly learning contrast rather than shared semantics.

### 4.2 Effect of the Memory Module

Since the memory module serves as a key component that records essential multimodel features and the pivot receiving optimized signal from RCR, it is of great importance to investigate its effect on model performance. Specifically, we explore the effect of the number of memory vectors $N$ and run LLM-based models on HMC and Memeplate datasets[9]. The performance (y-axis) of models with respect to the value of $N$ (x-axis) is illustrated in Figure 3. There are several observations. First, for both datasets, when the value of $N$ is small, increasing its value brings significant enhancement to model performance. This observation is intuitive in that more memory vectors provide a larger parameter space to comprehensively accommodate enough information between multiple modalities and thus lead to better performance. Second, when the value of $N$ is high, the performance improvement brought by the increase of $N$ is moderate. This indicates that when the number of memory vectors reaches a certain point, no more useful contrast information for the task is leveraged and thus results in less improvements.

In addition, we investigate the effect of the memory module when it works with RCR by replacing the memory module with other widely used architectures, namely, outer product operation (OP) and co-attention (Co-Att) (Lu et al., 2016). Specifically, for OP, we firstly obtain the visual feature $\mathbf{v}$ and text features $\mathbf{t}$ using the same process as our approach. Then, we compute the outer product of $\mathbf{v}$ and $\mathbf{t}$, and flatten the resulting matrix into a vector to represent the contrast vector $\mathbf{m}_x$. For Co-Att, we apply co-attention to fuse $\mathbf{v}$ and $\mathbf{t}$ and regard the output as $\mathbf{m}_x$. The results for different datasets are reported in Table 7. It is observed that their performance is worse than the performance of our approach with memory, which confirms the effectiveness of our approach in leveraging the memory and RCR for multimodal classification tasks that require contrast information modeling.

### 4.3 Case Study

We also investigate three similar memes for qualitative analysis for multimodal sentiment analysis. The images and texts with their predictions from different models, as well as the gold standard, are illustrated in Figure 4, where (a) and (b) have the same texts; (a) and (c) use the same image. As

[9]We select the two representative datasets for different languages and tasks, one for English HMD and the other for Chinese humor recognition.

| | HMC | | MEMOTION7K | | | MULTIOFF | | MEMEPLATE | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | AUROC | T1-F1 | T2-F1 | T3-F1 | ACC | F1 | ACC | F1 |
| **OP** | $76.44_{\pm 0.18}$ | $84.86_{\pm 0.22}$ | $39.42_{\pm 0.22}$ | $50.67_{\pm 0.21}$ | $34.68_{\pm 0.22}$ | $71.97_{\pm 0.20}$ | $65.90_{\pm 0.18}$ | $55.11_{\pm 0.18}$ | $48.62_{\pm 0.22}$ |
| **CO-ATT** | $76.63_{\pm 0.21}$ | $84.98_{\pm 0.22}$ | $39.67_{\pm 0.20}$ | $50.93_{\pm 0.22}$ | $34.77_{\pm 0.21}$ | $72.20_{\pm 0.22}$ | $66.12_{\pm 0.24}$ | $55.21_{\pm 0.20}$ | $48.89_{\pm 0.22}$ |
| **M** | $^{*}\mathbf{77.88}_{\pm 0.24}$ | $^{*}\mathbf{86.34}_{\pm 0.23}$ | $^{*}\mathbf{41.56}_{\pm 0.21}$ | $^{*}\mathbf{52.73}_{\pm 0.21}$ | $^{*}\mathbf{35.88}_{\pm 0.24}$ | $^{*}\mathbf{74.09}_{\pm 0.20}$ | $^{*}\mathbf{68.43}_{\pm 0.21}$ | $^{*}\mathbf{56.52}_{\pm 0.20}$ | $^{*}\mathbf{50.21}_{\pm 0.21}$ |

Table 7: Experiment results of different models using LLMs and RCR, where the memory module in our approach is replaced by two widely used approaches for multimodal feature fusion, namely, outer product operation (OP) and Co-attention mechanism (Co-Att). The performance of memory (M) with RCR is also presented for reference.


(a)

| | |
|---|---|
| Base: | *neural* |
| +M: | *neural* |
| +RCR: | *negative* |
| +M+RCR: | *negative* |
| Gold: | *negative* |

(b)

| | |
|---|---|
| Base: | *positive* |
| +M: | *positive* |
| +RCR : | *positive* |
| +M+RCR: | *positive* |
| Gold: | *positive* |

(c)

| | |
|---|---|
| Base: | *positive* |
| +M: | *positive* |
| +RCR : | *negative* |
| +M+RCR: | *positive* |
| Gold: | *positive* |

Figure 4: Demonstration of three memes for multimodal sentiment analysis with the polarities predicted by different models. The gold standard sentiment labels of all memes are also presented. The model produced labels that match the gold standard and the labels that do not match are highlighted in green and red colors, respectively.

a result, (a) conveys a negative sentiment polarity, while (b) and (c) have positive polarities. The predictions that match and do not match the gold standard are highlighted in green and red colors, respectively. By investigating the results, we observe that the three baselines struggle to predict sentiment labels that match the gold standards for all memes, whereas our approach is able to accurately identify sentiment polarities of all memes. A possible reason is the following. Negative sentiment polarities are generally derived from the contrast between multiple modalities. The baselines have limitations that prevent them from learning such contrast, either lacking a particular mechanism to do so or being equipped without effective guidance. In contrast, RCR or Memory+RCR provide enough information to learn such contrast and thus help our approach to correctly analyze its sentiment polarity.

## 5 Related Work

Recent studies for image-text understanding generally utilize advanced pre-trained visual and text encoders, such as ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), MMBT (Kiela et al., 2019), FLAVA (Singh et al., 2022), Flamingo (Alayrac et al., 2022), SLIP (Mu et al., 2022), BLIP2 (Li et al., 2023), etc. With LLMs demonstrating strong language modeling capabilities and achieving state-of-the-art results on many NLP tasks (Qin et al., 2021; Song et al., 2021; Achiam et al., 2023; Touvron et al., 2023a,b; Taori et al., 2023; Chiang et al., 2023; Tian et al., 2023b, 2024), there are studies that combines visual encoder and LLMs, such as MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023). These studies mainly train the models to align image and text features, focusing on image-text matching and achieving satisfying performance accordingly (Park and Paik, 2023; Ramos et al., 2023; Wang et al., 2023a; Ma et al., 2024). Although there are some studies modeling the contrast between multiple modalities (Radford et al., 2021; Lee et al., 2022; Wang et al., 2024), their focus is still on the alignment of image and text features. However, besides image-text matching, there are multimodal tasks that rely on the contrast rather

than the shared content between multiple modalities, such as HMD, multimodal sentiment analysis, offensive content detection, and humor recognition (Suryawanshi et al., 2020; Sharma et al., 2020; Pramanick et al., 2021a,b; Kocoń et al., 2021; Sharma et al., 2022a,b; Hakimov et al., 2022). To perform these tasks, most existing studies still follow the paradigm of simply aligning and fusing image and text features with a particular module or operation, such as vector concatenation, attentions, and contrastive learning (Goyal et al., 2022; Xu et al., 2022; Liang et al., 2022; Pramanick et al., 2022; Kumar and Nandakumar, 2022; Hee et al., 2023; Qu et al., 2023; Wang et al., 2023b; Ayetiran and Özgöbek, 2023; Kumari et al., 2023). To further enhance these tasks, there are studies that ensembles models to benefit from the outputs of different models from various aspects (Lippe et al., 2020; Sandulescu, 2020; Muennighoff, 2020) or utilize additional training data and features to enhance the capability of models to capture multimodal information (Velioglu and Rose, 2020; Zhu, 2020).

Compared with existing studies, our approach differs from them by explicitly modeling the contrast between multiple modalities rather than modeling how well images and texts are aligned. Particularly, we design a memory module with RCR to learn the contrast, where a reward model is trained to assess how well the memory module learns the contrast and an effective learning approach with self-rejection sampling is applied, which, to our best knowledge, is not used before in previous studies for similar tasks.

## 6 Conclusion

In this paper, we propose an LLM-driven approach for learning contrast with cross-modal memory and RCR, which learns and enhances the contrast information between visual and text features that helps final classification results. We perform several tasks that require modeling multimodal contrast, including HMD, multimodal sentiment analysis, offensive content detection, and humor recognition, etc. Experimental results on English and Chinese benchmark datasets confirm the validity of the proposed approach, which outperforms strong baselines and existing studies and achieves state-of-the-art performance. Further analysis also confirms that the combination of memory and RCR demonstrates their superiority in learning contrast between multimodalities and thus facilitating downstream tasks.

## Acknowledgements

This paper is supported by the National Key Research and Development Program of China under the grant (2023YFC3303800).

## References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a Visual Language Model for Few-shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Eniafe Festus Ayetiran and Özlem Özgöbek. 2023. An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. *Hate Speech and Offensive Language Detection*.

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

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023a. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023b. Prompting for Multimodal Hateful Meme Classification. *arXiv preprint arXiv:2302.04156*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. *GitHub Repository*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob

Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*, pages 1–21.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision Models are More Robust and Fair when Pretrained on Uncurated Images without Supervision. *arXiv preprint arXiv:2202.08360*.

Yingmei Guo, Jinfa Huang, Yanlong Dong, and Mingxing Xu. 2020. Guoym at SemEval-2020 Task 8: Ensemble-based Classification of Visuo-Lingual Metaphor in Memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1120–1125, Barcelona (online).

Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. 2022. TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes. *arXiv preprint arXiv:2204.06299*.

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*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. *arXiv preprint arXiv:2308.12038*.

Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. IITK at SemEval-2020 Task 8: Unimodal and Bimodal Sentiment Analysis of Internet Memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1135–1140, Barcelona (online).

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, 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. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *Advances in neural information processing systems*, 33:2611–2624.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, Aggressive, and Hate Speech Analysis: From Data-centric to Human-centered Approach. *Information Processing & Management*, 58(5):102643.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. MemeFier: Dual-stage Modality Fusion for Image Meme Classification. *arXiv preprint arXiv:2304.02906*.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. *NLP4PI 2022*, page 171.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. EmoffMeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, pages 1–36.

Gitanjali Kumari, Amitava Das, and Asif Ekbal. 2021. Co-attention based Multimodal Factorized Bilinear Pooling for Internet Memes Analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India.

Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. 2022. Uniclip: Unified framework for contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, 35:1008–1019.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling Hate in Online Memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-efficient Prompt Tuning. *arXiv preprint arXiv:2104.08691*.

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*.

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. *arXiv preprint arXiv:1908.03557*.

Zefeng Li, Hongfei Lin, Liang Yang, Bo Xu, and Shaowu Zhang. 2022. Memeplate: A Chinese Multimodal Dataset for Humor Understanding In Meme Templates. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, page 527–538, Berlin, Heidelberg.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multimodal Contrastive Representation Learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

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*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt Tuning Can be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-agnostic Visiolinguistic Representations for Vision-and-language Tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-image Co-attention for Visual Question Answering. *Advances in neural information processing systems*, 29.

Ziyu Ma, Shutao Li, Bin Sun, Jianfei Cai, Zuxiang Long, and Fuyan Ma. 2024. GeReA: Question-Aware Prompt Captions for Knowledge-based Visual Question Answering. *arXiv preprint arXiv:2402.02503*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. SLIP: Self-supervision Meets Language-image Pre-training. In *European Conference on Computer Vision*, pages 529–544.

Niklas Muennighoff. 2020. Vilio: State-of-the-art Visiolinguistic Models Applied to Hateful Memes. *arXiv preprint arXiv:2012.07788*.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named Entity Recognition for Social Media Texts with Semantic Augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online.

Sofiane Ouaari, Tsegaye Misikir Tashu, and Tomáš Horváth. 2022. Multimodal feature extraction for memes sentiment classification. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 285–290.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Seokmok Park and Joonki Paik. 2023. RefCap: image captioning with referent objects attributes. *Scientific Reports*, 13(1):21577.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic.

Han Qin and Yan Song. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland.

Han Qin, Yuanhe Tian, and Yan Song. 2021. Relation Extraction with Word Graphs from N-grams. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2860–2868, Online and Punta Cana, Dominican Republic.

Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 293–310.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*.

Vlad Sandulescu. 2020. Detecting Hateful Memes Using a Multimodal Deep Ensemble. *arXiv preprint arXiv:2012.13235*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online).

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: Detecting the Victims Targeted by Harmful Memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022b. Findings of the CONSTRAINT 2022 Shared Task on Detecting the Hero, the Villain, and the Victim in Memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language and Vision Alignment Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Yan Song, Shuming Shi, and Jing Li. 2018. Joint Learning Embeddings for Chinese Words and Their Components via Ladder Structured Networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4375–4381.

Yan Song, Tong Zhang, Yonggang Wang, and Kai-Fu Lee. 2021. ZEN 2.0: Continue Training and Adaption for N-gram Enhanced Text Encoders. *arXiv preprint arXiv:2105.01279*.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. *GitHub repository*.

Yuanhe Tian, Weidong Chen, Bo Hu, Yan Song, and Fei Xia. 2023a. End-to-end Aspect-based Sentiment Analysis with Combinatory Categorial Grammar. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13597–13609, Toronto, Canada.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023b. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *arXiv preprint arXiv:2311.06025*.

Yuanhe Tian, Yan Song, and Fei Xia. 2022. Improving Relation Extraction through Syntax-induced Pretraining with Dependency Masking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue Summarization with Mixture of Experts based on Large Language Models. In *The 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. LLaMA 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Tsinghua University. 2023. VisualGLM-6B. *GitHub repository*.

Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning Solution to Hateful Memes Challenge. *arXiv preprint arXiv:2012.12975*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, Costin Chiru, and Stefan Trausan-Matu. 2020. UPB at SemEval-2020 Task 8: Joint Textual and Visual Modeling in a Multi-Task Learning Architecture for Memotion Analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1208–1214, Barcelona (online).

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. 2024. COSMO: COntrastive Streamlined MultimOdal Model with Interleaved Pre-Training. *arXiv preprint arXiv:2401.00849*.

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. 2023a. Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2617–2625.

Qianlong Wang, Hongling Xu, Zhiyuan Wen, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2023b. Image-to-Text Conversion and Aspect-Oriented Filtration for Multimodal Aspect-Based Sentiment Analysis. *IEEE Transactions on Affective Computing*.

Haojie Xu, Weifeng Liu, Jiangwei Liu, Mingzheng Li, Yu Feng, Yasi Peng, Yunwei Shi, Xiao Sun, and Meng Wang. 2022. Hybrid Multimodal Fusion for Humor Detection. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 15–21.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335*.

Ziqing Yang, Yuchen Pan, and Yiming Cui. 2023. Visual-Chinese-LLaMA-Alpaca. *GitHub repository*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022a. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Yazhou Zhang, Lu Rong, Xiang Li, and Rui Chen. 2022b. Multi-modal Sentiment and Emotion Joint Analysis with a Deep Attentive Multi-Task Learning Model. In *European Conference on Information Retrieval*, pages 518–532.

Qi Zhong, Qian Wang, and Ji Liu. 2022. Combining knowledge and multi-modal fusion for meme classification. In *International Conference on Multimedia Modeling*, pages 599–611.

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. *arXiv preprint arXiv:2012.08290*.