MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification

Siddhant Bikram Shah[1] **Shuvam Shiwakoti**[2] **Maheep Chaudhary**[3] **Haohan Wang**[4]
[1]Northeastern University, USA
[2]Delhi Technological University, India
[3]Nanyang Technological University, Singapore
[4]University of Illinois Urbana-Champaign, USA

Abstract

The complexity of text-embedded images presents a formidable challenge in machine learning given the need for multimodal understanding of multiple aspects of expression conveyed by them. While previous research in multimodal analysis has primarily focused on singular aspects such as hate speech and its subclasses, this study expands this focus to encompass multiple aspects of linguistics: hate, targets of hate, stance, and humor. We introduce a novel dataset PrideMM comprising 5,063 text-embedded images associated with the LGBTQ+ Pride movement, thereby addressing a serious gap in existing resources. We conduct extensive experimentation on PrideMM by using unimodal and multimodal baseline methods to establish benchmarks for each task. Additionally, we propose a novel framework MemeCLIP for efficient downstream learning while preserving the knowledge of the pre-trained CLIP model. The results of our experiments show that MemeCLIP achieves superior performance compared to previously proposed frameworks on two real-world datasets. We further compare the performance of MemeCLIP and zero-shot GPT-4 on the hate classification task. Finally, we discuss the shortcomings of our model by qualitatively analyzing misclassified samples. Our code and dataset are publicly available at: https://github.com/SiddhantBikram/MemeCLIP.

1 Introduction

In recent years, the pervasive integration of social media platforms into everyday life has resulted in an exponential increase in the generation and dissemination of multimedia content. At the heart of this digital ecosystem lies the meme: a text-embedded image imbued with humor, wit, and often, a subversive edge, which offers a medium through which individuals can express opinions, share experiences, and engage in online activism (Moreno-Almeida, 2021; Baker et al., 2020). With their ability to distill complex ideas into digestible units of communication, memes have emerged as a powerful medium for expressing both support and opposition toward socio-political events (Imperato et al., 2023).

However, with opinions being expressed freely, hate speech becomes prevalent, often directed towards individuals, organizations, and even marginalized communities (Thapa et al., 2023), targeting them with vitriol and prejudice (Lingiardi et al., 2020; Imperato et al., 2023). Particularly, the LGBTQ+ movement stands as a prominent subject of online discourse, where memes serve as vehicles of both solidarity and resistance, reflecting the multifaceted dynamics of attitudes and perceptions within the community and beyond (Gal et al., 2016). In this context, the distinction between humor and harm becomes blurred, as memes straddle the line between satire and offense, challenging researchers and platforms alike to navigate the complexities of online content moderation (Langvardt, 2017). Previous attempts that endeavored to suppress such content have resulted in the discriminative suppression of all LGBTQ+ content (Griffin, 2022, 2024), which can harm the awareness and acceptance of this community. Thus, understanding the nuances of hate speech, opinions, and intended humor within memes becomes paramount for fostering an inclusive digital environment and combating online discrimination.

To address these challenges, we introduce PrideMM: a novel dataset comprising 5,063 text-embedded images related to the LGBTQ+ movement annotated with a multi-aspect schema encompassing four tasks:

- **Task A**: Detection of Hate Speech
- **Task B**: Classifying the Targets of Hate Speech
- **Task C**: Classification of Topical Stance
- **Task D**: Detection of Intended Humor

The analysis of text-embedded images is partic-

(a) {0, NA, 1, 0} (b) {1, 1, 1, 1} (c) {1, 0, 2, 1} (d) {1, 3, 0, 1} (e) {1, 0, 1, 1}

Figure 1: Samples of text-embedded images from the PrideMM dataset annotated across four aspect labels. The labels are in the form of {Hate, Target, Stance, Humor}. For Hate, {0, 1} correspond to *No Hate* and *Hate* respectively. For Target, {0, 1, 2, 3} correspond to hate targeted towards *Undirected*, *Individual*, *Community*, and *Organization* respectively. For Stance, {0, 1, 2} correspond to *Neutral*, *Support*, and *Oppose* respectively. For Humor, {0, 1} correspond to *No Humor* and *Humor* respectively.

ularly challenging given the need for contextual understanding and the prevalence of ambiguity and subjectivity in them (Sherratt, 2022). Accordingly, the multi-aspect nature of our dataset provides a more holistic view of the diverse themes usually expressed through memes. Through PrideMM, we aim to cultivate a more profound understanding of interactions on social media through memes and facilitate the development of multimodal content moderation methods to make the internet a safer space. We implement a range of baseline and state-of-the-art hate speech detection models to establish benchmarks for each task of PrideMM. Sample images from PrideMM are illustrated in Figure 1 alongside their annotation labels.

We further propose MemeCLIP, a novel framework that leverages the knowledge of the Contrastive Language-Image Pre-Training (CLIP) model (Radford et al., 2021) by using multiple lightweight modules for multimodal and multi-aspect meme classification. We employ linear layers to effectively disentangle image and text representations in CLIP's multimodal embedding space. We utilize Feature Adapters to preserve the prior knowledge of CLIP and adapt its embedding spaces to the meme classification task while avoiding overfitting on smaller datasets. We further implement a cosine classifier alongside Semantic-Aware initialization (Shi et al., 2023) to make it more robust to the class imbalances that may exist in datasets such as PrideMM and HarMeme (Pramanick et al., 2021a) that are representative of real-world data distributions. Distinct from previous multimodal meme classification frameworks, MemeCLIP is trained end-to-end in a single step and does not rely on extraneous models to create augmented data. Our main contributions can be summarized as follows:

- We release PrideMM, a dataset containing 5,063 text-embedded images related to the LGBTQ+ movement.
- We benchmark PrideMM by using various unimodal and multimodal methods including existing multimodal frameworks proposed for meme classification.
- We introduce MemeCLIP, a novel framework that utilizes lightweight modules on top of a frozen CLIP model to classify memes.

2 Related Work

2.1 Multimodal Datasets

Multimodal image-text analysis has seen significant strides in recent years owing to the widespread popularity and availability of image-text pairs across social media. With the increasing need for hate speech and offensive content detection, multimodal datasets for hate speech detection have seen a particular surge. One of the first datasets in this domain was the Hateful Meme Challenge (HMC) dataset (Kiela et al., 2020), containing synthetic memes designed to convey contrastive implications from the image and text modalities that target religion, race, disability, and sex. Similarly, the Harm-C (Pramanick et al., 2021a) and Harm-P (Pramanick et al., 2021b) datasets comprise memes related to the COVID-19 pandemic and US politics respectively that were annotated across three degrees of harmfulness and four subclasses of hate speech targets. Bhandari et al. (2023) annotated samples for hate speech detection and target classification similarly, collecting text-embedded images related to the Russia-Ukraine conflict from Twitter, Facebook

Work	Data Source	Multimodal	Sub-Classes	Multi-aspect	Size	Context
Qu et al. (2022)	Reddit	✓	✗	✗	1,170	COVID-19, BLM, Veganism
Tanaka et al. (2022)	Meme Websites	✓	✗	✗	7,500	General Discourse
Kiela et al. (2020)	Self-Generated	✓	✗	✗	10,000	General Discourse
Suryawanshi et al. (2020)	FB, Twitter, Instagram	✓	✗	✗	743	U.S. Election
Pramanick et al. (2021a)	Google Images	✓	✓	✗	3,544	COVID-19
Pramanick et al. (2021b)	Google Images	✓	✓	✗	3,522	U.S. Politics
Bhandari et al. (2023)	Twitter, FB, Reddit	✓	✓	✗	4,723	Russia-Ukraine War
Dacon et al. (2022)	Reddit	✗	✓	✓	9930	LGBTQ+ Movement
Gautam et al. (2020)	Twitter	✗	✗	✓	9937	#MeToo Movement
Ousidhoum et al. (2019)	Twitter	✗	✓	✓	13,000	General discourse
PrideMM (Ours)	FB, Twitter, Reddit	✓	✓	✓	5,063	LGBTQ+ Movement

Table 1: Summary of datasets used in the literature.

(FB), and Reddit. Tangentially, Suryawanshi et al. (2020) employed extensive annotation guidelines to create the MultiOFF dataset for offensive content detection, consisting of memes collected from Reddit, Facebook, Twitter, and Instagram. In an effort to discern humor often expressed in memes, Tanaka et al. (2022) created a humor detection dataset by proposing a pipeline to extract memes devoid of interpersonal influence on the perception of humor. To identify disinformative memes, Qu et al. (2022) introduced the DisinfoMeme dataset that contains memes related to COVID-19, the Black Lives Matter (BLM) movement, and Veganism.

2.2 Multi-aspect Datasets

Online discourse on socio-political events is often imbued with a series of human emotions, leading researchers to study the numerous aspects of linguistics expressed in them. Dacon et al. (2022) used comments collected from RedditBias (Barikeri et al., 2021) related to LGBTQ+ individuals and annotated each comment for the presence of Toxicity, Severe Toxicity, Obscene, Threat, Insults, and Identity Attacks. Similarly, Gautam et al. (2020) curated a dataset of tweets related to the #MeToo movement in social media by annotating the tweets across five different aspects. Taking multi-aspect datasets one step further, Ousidhoum et al. (2019) compiled an extensive multi-aspect Twitter dataset with English, French, and Arabic samples, with each annotated for different aspects including hate and offensiveness. Table 1 provides a detailed comparison of the datasets cited in this section.

Multi-aspect data helps better encompass the spectrum of human emotions that may be associated with social media interactions. Most multimodal datasets, while only focusing on a single aspect and its sub-classes, fail to encompass the complex dynamics of emotions expressed by the masses. Our work aims to address this gap by presenting a multimodal and multi-aspect dataset comprising three different aspects- hate, topical stance, and humor, and one subclass within hate: targets of hateful speech, to enable more nuanced studies of multimodal meme data through computational methods.

2.3 Multimodal Frameworks

Recent developments in large vision-language models have incited a wave of research in methods to tackle hate speech in text-embedded images. MOMENTA (Pramanick et al., 2021b) was one of the first frameworks proposed to incorporate CLIP's vision and language encoders for multimodal hate speech classification. It extracts regions of interest from image data and named entities from text data to combine them with CLIP representations by using cross-modal attention fusion. Similarly, HateCLIPper (Kumar and Nandakumar, 2022) was proposed to better model cross-modal interactions between CLIP representations. Textual inversion (Gal et al., 2022) has been used to integrate visual cues in the text representation space in frameworks such as ISSUES (Burbi et al., 2023). Recent works make use of image caption models to extract text captions from images and learn a single language processing model (Cao et al., 2023b,a). However, rather than relying on augmented data from extraneous models, our proposed framework MemeCLIP leverages the knowledge learned by CLIP's encoders during its pre-training step to process the rich multimodal information inside each image. Additionally, we use Feature Adapters alongside residual connections to prevent overfitting as annotated datasets for multimodal meme classification generally lack a high number of samples. We further utilize a cosine classifier to make MemeCLIP more robust to imbalanced data classes, which is prevalent in multi-label tasks in this domain.

3 Dataset

In this section, we describe various aspects of our dataset including data collection, annotation guidelines, and dataset statistics. Our dataset comprises 5,063 text-embedded images that encompass memes, posters, and infographics relevant to the LGBTQ+ movement. We only include images from 2020-2024 as this period saw an upsurge of social media content in this domain (Oz et al., 2023). This also allows our dataset to represent contemporary social media interactions through memes. Note that by the term LGBTQ+, we refer to all gender identities and sexual orientations inclusively.

3.1 Data Collection

To maintain diversity in the dataset, we collected data from three popular social media platforms: Facebook, Twitter, and Reddit, through manual search and extraction. For Twitter, we used hashtags such as *#lgbt*, *#pride*, *#trans*, *#transrights*, *#nonbinary*, and *#genderidentity* to filter images related to LGBTQ+ discussions. For Facebook, we targeted groups that frequently discussed LGBTQ+ content. Similarly, for Reddit, we identified subreddits where discussion related to LGBTQ+ was more prominent. Further, to ensure the relevance and quality of the dataset, the data collection process was subject to filtering criteria. Detailed filtering criteria for our dataset can be found in Appendix A.2. As different annotators may encounter and collect the same image, we sequentially employed two image deduplication tools: dupeGuru[1] and difPy[2], to search for duplicates and retain the highest quality image out of each batch of duplicates. We used the OCR application provided by Google Cloud Vision API[3] to extract textual data from the images. We removed non-alphanumeric elements such as special characters, hyperlinks, symbols, and non-English characters to reduce noisy text data and ensure data quality. Note that the text may occasionally contain unintentional noisy artifacts.

3.2 Data Annotation

We engaged five experienced annotators well-versed in NLP and computational linguistics to annotate data samples for PrideMM. The annotators had a prior understanding of the LGBTQ+ movement and meme archetypes on social media. We presented them with comprehensive annotation guidelines to ensure uniform and unbiased annotations, and asked them to annotate each image separately for all four tasks. A 3-phase annotation schema was used to ensure accurate and consistent annotations. First, a dry run was conducted to evaluate the understanding of the annotation guidelines among the annotators where every annotator was given an identical batch of 50 images for annotation. Second, a revision phase was conducted where every annotator was given another identical batch of 200 images and received a revised set of instructions based on the results of the first phase. Finally, in the consolidation phase, the annotators annotated a final batch of 50 images while discussing and revising the annotation guidelines until a consensus was reached. These steps were taken to minimize misannotations and noisy labels in the PrideMM dataset. The meticulously devised annotation guidelines were followed to ensure consistency in the annotations. Each image in our dataset was independently annotated for the three aspects and one sub-class, apart from the connection between 'Hate' and 'Hate Targets'.

3.3 Annotation Guidelines

In this section, we describe the annotation guidelines used to annotate the dataset. We devise separate guidelines for each of the four tasks.

Hate Speech. This task aimed to identify instances of hate speech in the images. The primary focus was on identifying images that intentionally conveyed hateful sentiments. Annotators needed to distinguish between images expressing strong disagreement without resorting to offensive language and those containing genuine elements of hate speech. This differentiation aimed to guarantee accurate labeling, ensuring that images conveying genuinely hateful sentiment through visual content, language, or a combination of both were appropriately identified.

Hate Targets. This task required annotators to identify the targets of hate in hateful images by classifying the images into one of the four classes: *Undirected*, *Individual*, *Community*, and *Organization*. Images were labeled as *Undirected* when they targeted abstract topics, societal themes, or ambiguous targets like 'you' that were not directed toward any specific individuals, entities, or groups. Hateful images targeting specific people including political leaders, celebrities, or activists like 'Joe

[1] https://github.com/arsenetar/dupeguru

[2] https://github.com/elisemercury/Duplicate-Image-Finder

[3] https://cloud.google.com/vision/docs

Biden' and 'J.K. Rowling' were annotated as *Individual*. Likewise, the label *Community* was used for instances of images targeting broader social, ethnic, or cultural groups like 'LGBT' or 'trans'. Lastly, images targeting corporate entities, institutions, or similar organizations like 'Chick-fil-A' and 'government" were annotated as *Organization*.

Stance. This task involved annotating the images into either of three distinct categories: *Support*, *Oppose*, and *Neutral*, determined by their stance within the context of the LGBTQ+ movement. The *Support* label was given to images that expressed support towards the goals of the movement, agreed with efforts in fostering equal rights for LGBTQ+ individuals, and promoted awareness for the movement's goals. The *Oppose* label was given to images that conveyed disagreement with the goals of the movement, denied the problems faced by individuals who identified as LGBTQ+, and dismissed the need for equal rights and acceptance. The *Neutral* label was given to images that were contextually relevant to the movement but did not exhibit support or opposition towards the movement.

Humor. In this task, annotators were asked to identify images showcasing humor, sarcasm, or satire related to the LGBTQ+ Pride movement. Annotators were instructed to discern the presence of humor in the images regardless of whether they presented a lighthearted or insensitive perspective on serious subjects. Note that annotators were asked to annotate images based on whether the creator of the image intended for it to be humorous, and not based on whether the annotator personally found it humorous. This task aimed to capture the nuanced use of text-embedded images for comedic or satirical purposes, thereby helping disentangle hate and humor in the images related to this movement.

3.4 Statistics and Inter-Annotator Agreement

Table 2 shows the distribution of images in PrideMM across all class labels. For the hate detection task, the dataset has a balanced distribution of binary labels. The target classification task exhibits a heavily imbalanced distribution. Given the context of this study, most hateful images convey undirected hate or are targeted toward communities, with a low frequency of hate against individuals and organizations. For the stance classification task, the number of images is well-balanced across three labels. On the other hand, as memes are often meant to be humorous, the majority of the images in the dataset are annotated to humor. We use topic modeling to analyze PrideMM's text content, and the results are presented in Appendix A.1.

We used the Fleiss' Kappa (κ) (Falotico and Quatto, 2015) as a statistical measure to assess the inter-annotator agreement across all four tasks. For Task A (Hate Speech detection), κ was 0.66/0.74 in the dry run and final phase respectively, for Task B (Target detection), κ was 0.68/0.81, for Task C (Stance detection), κ was 0.62/0.75, and for Task D (Humor detection), κ was 0.60/0.74. The increase in κ from the dry run phase to the final phase across all tasks reflects the effectiveness of the 3-phase annotation schema.

Task	Label	#Samples	%
Hate	No Hate	2,581	50.97%
	Hate	2,482	49.03%
Target	Undirected	771	31.07%
	Individual	249	10.03%
	Community	1,164	46.90%
	Organization	298	12.00%
Stance	Neutral	1,458	28.80%
	Support	1,909	37.70%
	Oppose	1,696	33.50%
Humor	No Humor	1,642	32.43%
	Humor	3,421	67.57%

Table 2: Dataset Statistics for PrideMM. The data consists of 5,063 samples for Hate, Stance, and Humor classification tasks, and 2,482 samples for the Target classification task.

4 Methodology

In this section, we describe our proposed framework, MemeCLIP, for multi-aspect meme classification. We utilize the vision-language model CLIP (Radford et al., 2021) to create rich representations that effectively encapsulate the semantics of a meme. We add lightweight modules on top of CLIP to disentangle image and text representations, prevent overfitting, and make MemeCLIP more robust to imbalanced data. Figure 2 illustrates the overall architecture of MemeCLIP. Below, we describe each component of MemeCLIP in detail.

Zero-shot CLIP. The vision-language model CLIP exhibits stellar zero-shot performance and transfer learning capabilities (Radford et al., 2021; Wang et al., 2023). CLIP is pre-trained on 400 million image-text pairs from the internet, enabling it to encode visual and textual data in a shared embedding space. The model consists of an Im-

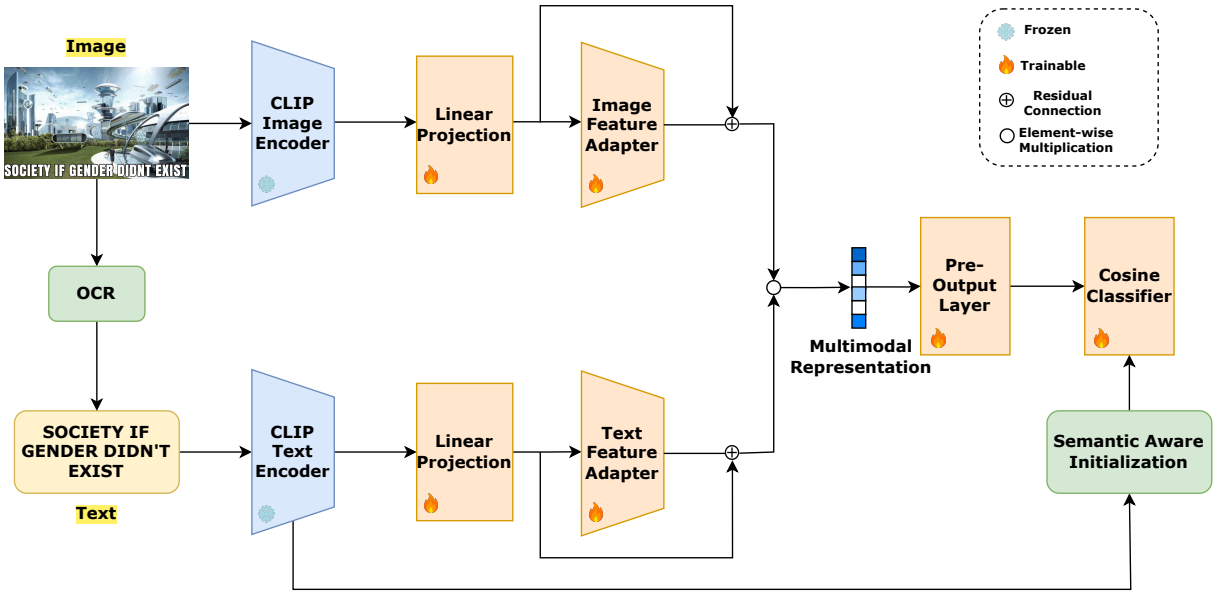

Figure 2: An overview of our proposed framework, MemeCLIP. We use frozen CLIP image and text encoders to create representations for each image-text pair. These representations are passed through linear layers to disentangle the modalities in CLIP's shared embedding space. We implement Feature Adapters with residual connections for each modality to prevent overfitting. We use a cosine classifier to make MemeCLIP more robust to imbalanced data. We initialize classifier weights by using Semantic-Aware Initialization to further improve performance.

age Encoder E_I and a Text Encoder E_T. We freeze the weights of both encoders to preserve the valuable knowledge captured by them during pre-training. The unimodal image and text representations $F_I, F_T \in \mathbb{R}^{768}$ effectively encapsulate the semantics of a meme and are defined as:

$$F_I = E_I(I); \;\; F_T = E_T(T) \tag{1}$$

where I is the image and T is its text pair.

Linear Projection Layers. While the contrastive pre-training objective of CLIP promotes similarity between corresponding text and image pairs, memes often involve contrastive visual and linguistic content to evoke a sense of irony. Similar to (Kumar and Nandakumar, 2022), we employ individual linear projection layers for each modality to effectively disentangle image and text representations in the shared embedding space. These projection layers result in the unimodal projections $F_I^{proj}, F_T^{proj} \in \mathbb{R}^{1024}$, mapping the representations to the dimensions of CLIP's last hidden state, $D_{CLIP} \in \mathbb{R}^{1024}$, which enables the use of Semantic-Aware Initialization.

$$F_I^{proj} = L_I^{proj}(F_I); \;\; F_T^{proj} = L_T^{proj}(F_T) \tag{2}$$

Here, L_I^{proj} and L_T^{proj} represent the image and text projection layers respectively.

Feature Adapters. Since CLIP is pre-trained on an extensive amount of data, it may exhibit symptoms of overfitting when applied to smaller datasets for downstream tasks. Inspired by (Gao et al., 2024), we adopt lightweight Feature Adapters for both image and text modalities to learn the features of new data while retaining CLIP's prior knowledge. We further utilize residual connections to integrate prior image and text projections with the outputs of the Adapters, allowing our model to balance the knowledge of the fine-tuned adapter and the disentangled image and text projections. We use a residual ratio α to maintain harmony between these two modules. With the image and text Feature Adapters A_I and A_T respectively, the final unimodal representations $F_I, F_T \in \mathbb{R}^{1024}$ are obtained as follows:

$$F_I = \alpha A_I(F_I^{proj}) + (1-\alpha)F_I^{proj} \tag{3}$$

$$F_T = \alpha A_T(F_T^{proj}) + (1-\alpha)F_T^{proj} \tag{4}$$

Modality Fusion. Owing to the extensive unimodal feature modeling, MemeCLIP avoids the need for trainable fusion layers like Cross-Modal Attention Fusion in MOMENTA (Pramanick et al., 2021b) and Combiner in ISSUES (Burbi et al., 2023). We fuse the image and text representations by using an element-wise multiplication operation

(∘) to obtain a single multimodal representation $F_{MM} \in \mathbb{R}^{1024}$. We further alleviate the need for the two-stage training process employed in HateCLIPper (Kumar and Nandakumar, 2022) and ISSUES.

$$F_{MM} = F_I \circ F_T \tag{5}$$

F_{MM} is then passed through a linear pre-output layer before classification.

Classification. For classification, we employ a cosine classifier (Liu et al., 2020) that is robust to biases in prediction under class imbalances. Following (Shi et al., 2023), we adopt Semantic-Aware Initialization (SAI) to initialize the weights of this classifier by exploiting the semantic knowledge held within the text encoder of CLIP. We encode class labels by using the prompt "A photo of {LABEL}" into $F_{class} \in \mathbb{R}^{n \times 1024}$ where n is the number of classes. We use F_{class} to initialize the classifier weight W_{class}. During training, the predicted logit Z for a class x is calculated as follows:

$$Z_x = \sigma \times \frac{W_x \times F_{MM}}{||W_x||_2 \ ||F_{MM}||_2} \tag{6}$$

where σ is a static scaling factor for the cosine classifier.

5 Experimental Results

Table 3 and Table 4 show our experimental results for the PrideMM and HarMeme datasets respectively. We pre-define train/validation/test splits in the ratio 85/5/10 respectively for PrideMM, and use the pre-defined split for the HarMeme dataset. We conduct experiments on unimodal and multimodal baseline methods, and previous frameworks proposed for multimodal meme classification. We conduct each experiment on three random seeds and report the Mean and Standard Deviation (±) values for Accuracy, AUC (Macro), and F1-Score (Macro). We use ViT-L/14 as the image encoder for all CLIP-based methods except for MOMENTA, which uses ViT-B/32 as the backbone for its CLIP model by default. For CLIP, we use concatenation to fuse the unimodal feature representations. Further implementation details are outlined in A.3.

5.1 PrideMM Dataset

Unimodal Methods. For the unimodal methods, we used ViT-L/14 (Dosovitskiy et al., 2020) and CLIP's image encoder as image-based methods, and BERT (Devlin et al., 2018) and CLIP's text encoder as text-based methods. The image-based methods generally performed better than their text-based counterparts across all tasks, substantiating that Transformer-based visual models create meaningful representations that also capture the semantic meaning conveyed by the text embedded in the pixel space (Burbi et al., 2023). The text-based methods showed poor performance on the multi-label target and stance detection tasks. The image encoder of CLIP shows superior results compared to the standard pre-trained Visual Transformer while having the same architecture, demonstrating the effectiveness of contrastive pre-training. The unimodal methods generally perform worse than any multimodal method across all tasks, underscoring the need for multimodal processing in meme analysis.

Multimodal Methods. We tested the performance of the multimodal methods CLIP (Radford et al., 2021), CLIP-Adapter (Gao et al., 2024), MOMENTA (Pramanick et al., 2021b), HateCLIPper (Kumar and Nandakumar, 2022), ISSUES (Burbi et al., 2023), and our framework, MemeCLIP. MemeCLIP outperforms the baseline CLIP model and previously proposed multimodal methods across all metrics in the hate and humor classification tasks. Our model performs particularly well in the four-class target classification task, which has less than half the number of samples as the other tasks and harbors a heavy class imbalance, demonstrating the model's robustness to overfitting on majority classes in imbalanced datasets. Smaller frameworks with a lower number of parameters such as the baseline CLIP, CLIP-Adapter, HateCLIPper, and MemeCLIP perform optimally in this task with CLIP-Adapter showing the highest AUC, while the remaining methods show symptoms of overfitting. In the stance classification task, HateCLIPper surpasses MemeCLIP in accuracy, while the latter shows a higher AUC and F1-Score. While the CLIP-Adapter model is similar to MemeCLIP, it uses a single feature adapter as individual feature adapters for both modalities may carry redundant information from the unimodal encoders to the classifier; however, MemeCLIP outperforms this model in most tasks by using feature adapters for both image and text modalities, which may signify that separate layers for each modality are more effective at encoding memes that may contain visual and language content that convey different meanings individually and combined.

Method	Hate			Target			Stance			Humor		
	Acc.	AUC.	F1	Acc.	AUC.	F1	Acc.	AUC.	F1	Acc.	AUC.	F1
BERT	71.12±0.67	75.33±0.50	70.06±0.22	54.25±1.32	75.52±0.77	54.03±1.51	52.30±0.91	67.10±0.74	51.11±1.34	71.04±0.55	72.25±1.77	64.60±0.98
CLIP Text-Only	68.64±0.86	74.52±0.95	68.62±0.88	50.34±0.62	72.67±1.34	47.65±1.18	50.43±1.08	67.60±0.55	49.26±1.25	69.23±1.54	70.52±1.42	62.02±1.69
ViT-L/14	69.23±2.22	77.05±0.48	68.24±3.01	58.36±1.56	79.13±1.48	50.24±2.93	58.80±1.26	73.20±4.62	56.14±4.76	73.04±2.20	77.71±1.13	69.18±2.35
CLIP Img-Only	70.01±0.78	80.53±0.42	72.66±3.08	60.32±1.77	80.89±0.73	57.19±3.25	61.01±0.82	77.48±0.66	57.87±0.78	76.14±0.19	82.1±1.42	72.37±1.13
CLIP	72.39±1.20	80.47±0.61	72.33±1.26	61.14±0.59	81.92±0.44	58.46±1.02	59.31±0.82	76.92±0.87	57.81±1.14	76.66±1.32	80.73±0.20	73.23±1.56
CLIP-Adapter	72.75±1.09	80.91±0.56	72.69±1.02	61.59±0.52	**82.14**±0.35	58.08±0.91	59.55±0.47	77.23±0.73	57.93±0.91	77.01±1.01	80.97±0.71	73.51±0.97
MOMENTA	72.23±0.58	78.55±0.50	71.78±0.35	57.28±1.26	78.89±1.23	52.79±1.84	55.62±1.90	73.64±2.35	54.84±2.28	74.16±2.17	77.38±1.63	71.34±2.70
HateCLIPper	75.53±0.58	83.12±0.44	74.08±0.37	62.49±2.06	80.32±1.42	56.77±0.72	**63.24**±0.69	77.99±1.25	57.15±0.76	76.13±0.19	83.50±0.51	75.41±0.28
ISSUES	74.68±1.62	84.17±0.45	73.64±2.48	61.25±2.00	78.73±0.21	58.30±0.17	59.39±1.08	77.02±1.93	57.27±1.40	78.95±0.88	84.78±0.60	75.73±2.17
MemeCLIP	**76.06**±0.23	**84.52**±0.31	**75.09**±0.20	**66.12**±0.47	81.66±0.25	**58.65**±0.97	62.00±0.12	**80.11**±0.15	**57.98**±1.91	**80.27**±0.52	**85.59**±0.23	**77.21**±0.79

Table 3: Classification performance of methods on the PrideMM dataset. The results are in the form of *Mean* ± *Standard Deviation*. Performance is reported across three evaluation metrics: Accuracy, AUROC (Macro), and F1-Score (Macro). The best performance is highlighted in **bold**.

Method	Acc.	AUC.	F1
BERT	71.05±0.70	76.34±0.48	68.83±0.49
CLIP Text-Only	73.79±0.23	79.31±0.81	71.60±0.21
ViT-L/14	77.27±1.71	85.53±0.41	75.55±2.63
CLIP Img-Only	79.38±0.25	88.54±2.53	78.66±0.11
CLIP	81.36±0.81	87.27±0.67	80.30±0.95
CLIP-Adapter	82.21±0.73	87.53±0.61	80.89±0.87
MOMENTA	82.44±0.65	87.88±0.37	81.49±0.45
HateCLIPper	83.68±0.62	90.83±0.46	83.31±0.41
ISSUES	81.31±1.05	91.98±0.61	80.45±0.87
MemeCLIP	**84.72**±0.45	**92.07**±0.34	**83.74**±0.43

Table 4: Classification performance of methods on the HarMeme dataset. The results are in the form of *Mean* ± *Standard Deviation*. Performance is reported across three evaluation metrics: Accuracy, AUROC (Macro), and F1-Score (Macro). The best performance is highlighted in **bold**.

5.2 HarMeme Dataset

To test the generalizability of MemeCLIP across other meme datasets, we perform experiments on the HarMeme dataset (Pramanick et al., 2021a), which consists of real-world hateful memes shared on social media in the context of COVID-19. Similar to previous studies, we find that the performance of the multimodal models surpasses the unimodal models' performance by a decent margin. MemeCLIP outperforms the other multimodal frameworks in this dataset, demonstrating its effectiveness over the state-of-the-art baselines.

5.3 Ablation Study

We conduct a systematic ablation study to assess the contribution of each component of MemeCLIP towards its performance. The results of our ablation experiments are presented in Table 5. We start with the CLIP ViT-L/14 model and gradually integrate each external module of MemeCLIP. The rise in performance when the projection layers are applied signifies the importance of adapting CLIP's embedding spaces to our downstream task by disentangling the image and text representations. While the introduction of Feature Adapters initially leads to a temporary dip in F1-Score, it ultimately enables our model to produce more refined image and text representations due to the added learnable parameters. Replacing the linear classifier with a cosine classifier boosts performance by modulating weight updates with a static scaling factor. Finally, Semantic-aware initialization completes MemeCLIP by initializing classifier weights according to the semantic differences in class labels encoded by CLIP's text encoder, enhancing generalization further.

CLIP	PL	FA	CC	SAI	Acc.	AUC.	F1
✓					72.39±0.61	80.47±1.20	72.33±1.26
✓	✓				74.66±0.03	81.68±0.36	75.04±0.45
✓	✓	✓			75.33±0.17	83.44±0.01	74.77±0.88
✓	✓	✓	✓		75.78±0.32	84.35±0.05	74.92±0.38
✓	✓	✓	✓	✓	**76.06**±0.23	**84.52**±0.31	**75.09**±0.20

Table 5: Ablation experiments performed on MemeCLIP using the hate detection task of the PrideMM dataset. The results are in the form of *Mean* ± *Standard Deviation*. PL, FA, CC, and SAI denote Projection Layers, Feature Adapters, Cosine Classifier, and Semantic-Aware Initialization respectively. The last line represents the complete framework. The best results are highlighted in **bold**.

5.4 Comparison with GPT-4

Table 6 compares the performance of MemeCLIP against zero-shot GPT-4 (Achiam et al., 2023) through Microsoft Copilot[4] (accessed July 2024). We used the prompt *"Is this image hateful or not? Consider if the image and its text are hateful towards individuals, communities, organizations, or an undirected target. Also, consider the context of the entities represented in the image. Reply with only a number, 0 for no, and 1 for yes."* to manually test GPT-4's performance on the hate classification task for 100 images each from the PrideMM and HarMeme dataset's test set. We qualitatively found that while GPT-4 showed stellar zero-shot performance for hate classification, it tends to make cautious predictions by classifying non-hateful images as hateful or unsafe to the detriment of performance. This behavior may be caused by the stringent safety measures applied to commercial LLMs by LLM providers (Korbak et al., 2023; Bai et al., 2022).

Method	PrideMM			HarMeme		
	Acc.	AUC.	F1	Acc.	AUC.	F1
GPT-4	70.00	-	69.38	78.00	-	74.41
MemeCLIP	**73.00**	**79.06**	**72.54**	**85.00**	**92.58**	**83.02**

Table 6: Performance comparison between MemeCLIP and the GPT-4 model provided by Microsoft Copilot on hate classification for PrideMM and harm classification for HarMeme. The best results are highlighted in **bold**.

5.5 Misclassification Analysis

We present two examples of images misclassified by MemeCLIP in Figure 3. The meme presented in Figure 3a is hateful and opposes the values of the LGBTQ+ Pride movement under the guise of benign imagery and text, but is misclassified as non-hateful and neutral. Figure 3b shows a hateful meme mocking an individual, but MemeCLIP classifies it as hate against a community since the text mentions communal words such as "trans" and "women". Both memes were correctly classified as humorous by the model.

6 Conclusion

In this work, we release PrideMM, a multimodal and multi-aspect dataset comprising 5,063 memes related to the LGBTQ+ movement, addressing a serious gap in data resources in this domain. This dataset provides memes annotated across three aspects and one sub-class, allowing for greater flexibility in the establishment of ethical guidelines by social media policymakers. We further introduce MemeCLIP, a lightweight yet effective framework to harness CLIP's knowledge for multimodal meme classification while not being reliant on extraneous models to create augmented data for training.

With each module of MemeCLIP, we tried to address pervasive problems in the domain of hateful image classification. Specifically, we used a cosine classifier to counter class imbalances, which are common in multi-aspect datasets with many classes for each aspect. We utilize feature adapters to mitigate overfitting as datasets in this domain have a relatively small scale. We further use linear projection layers to dissociate the image and text modalities in the representations created by the pre-trained CLIP encoders as the pre-training dataset of CLIP majorly consists of image-text pairs that convey the same overall meaning, which may not be the case with creative or sarcastic memes. We show that CLIP, one of the most basic multimodal models, when combined with lightweight additional modules, can compete with or outperform models that require extraneous models to create augmented data, or even VLMs such as GPT-4. Our work is grounded on the importance of a data-centric approach to solving problems in this domain, rather than creating larger frameworks and incorporating extraneous data.

Our dataset endeavors to foster a deeper understanding of online interaction, community-building, and social change, whereas our framework is a step toward effective content moderation that helps create an inclusive, diverse, and equitable digital environment for all.

(a) Real:{1, 2, 2, 1}
Predction:{0, 2, 0, 1}

(b) Real:{1, 1, 2, 1}
Prediction:{1, 2, 2, 1}

Figure 3: Examples of memes misclassified by MemeCLIP across four tasks. The labels are in the form of {Hate, Target, Stance, Humor}. Label details are outlined in Figure 1.

[4]https://copilot.microsoft.com/

Ethical Considerations

User Privacy. Our dataset only comprises text-embedded images collected from publicly accessible web pages with no inclusion of user data. To the best of our knowledge, there are no copyright concerns associated with them. Since our OCR tool may have potentially scraped user identifiers such as Twitter usernames containing '@', we removed such data from the text during the pre-processing step.

Biases. Due to the size of the dataset and the number of annotation tasks, we acknowledge that some data samples may be misclassified, and not all annotators would agree on the same labels for a given sample. Further, due to the subjective nature of this topic, unintentional biases in the dataset's distribution and annotation may exist. We endeavored to minimize such occurrences by using comprehensive annotation guidelines and a three-phase annotation schema.

Potential Risks. PrideMM contains images conveying targeted hate towards specific individuals, communities, ethnic groups, and other entities. While we release our dataset to help robustify content moderation and foster safer spaces online, this data may potentially be used to spread hate and discrimination. We ask researchers to be aware that the inherent biases present in the dataset may negatively influence hate speech detection and moderation methods. Further, we release our multi-aspect dataset in order to give social media policymakers more freedom to censor specific types of harmful memes. However, we acknowledge that this may also lead to over-moderation and negatively affect social media users' freedom of expression.

Annotation. Five annotators were hired to annotate the images for our dataset. The annotators were Indian university students aged 20-25 and were familiar with the LGBTQ+ Pride movement. They were compensated fairly as per the standard local rate. Given the nature of this study, we acknowledged that the annotators may find the images disturbing or distressing. The annotators were given the option to opt out of the annotation process at any given time.

Reproducibility Statement. We provide implementation details and hyperparameter configurations for all the implemented models in Appendix A.3. The PrideMM dataset, source code for MemeCLIP, and MemeCLIP's pre-trained weights are publicly available at https://github.com/SiddhantBikram/MemeCLIP.

Environmental Impact. Leveraging hardware such as GPUs to train deep learning models is known to have a significant environmental footprint, primarily due to their high energy consumption resulting in carbon emissions. To mitigate the environmental impact of our research, we adopted a fine-tuning technique for pre-trained deep-learning models. This method allowed our models to generalize faster on new datasets, resulting in fewer computational resources consumed. Once these models are trained using GPUs, they can be loaded on relatively lightweight CPUs for inference purposes, attenuating potential environmental consequences.

Limitations

The proposed dataset PrideMM encompasses memes posted from 2020 to 2024, representing a snapshot of social media interactions within this specific period, which may fail to capture the dynamics of the LGBTQ+ movement on social media over an extended timeframe. Further, due to the subjective nature of the LGBTQ+ Movement, the size of the dataset, and the number of annotators, the annotation process is inherently prone to biases. While we include images from multiple sources in our dataset, we acknowledge the limited scope of our dataset compared to the vast number of social media platforms. Further, the intricate nature of memes may not be completely captured by our aspects, and more domain-specific aspects may be used to capture the context better. Due to the limited size and availability of labeled datasets in this domain, supervised frameworks such as MemeCLIP may be surpassed by unsupervised methods such as LLMs for content moderation as the capabilities and model size of LLMs progress. Finally, MemeCLIP and other models trained on PrideMM and similar datasets may exhibit biased predictions due to the presence of unintentionally biased data.

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*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional

ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

James E Baker, Kelly A Clancy, and Benjamin Clancy. 2020. Putin as gay icon? memes as a tactic in russian lgbt+ activism. *LGBTQ+ activism in Central and Eastern Europe: Resistance, representation and identity*, pages 209–233.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1993–2002.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2832–2836.

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*.

Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon, and Jiliang Tang. 2022. Detecting harmful online conversational content towards lgbtqia+ individuals. *arXiv preprint arXiv:2207.10032*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Noam Gal, Limor Shifman, and Zohar Kampf. 2016. "it gets better": Internet memes and the construction of collective identity. *New media & society*, 18(8):1698–1714.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

Rachel Griffin. 2022. The sanitised platform. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 13:36.

Rachel Griffin. 2024. The heteronormative male gaze: Experiences of sexual content moderation among queer instagram users in berlin. *International Journal of Communication*, 18:23.

Chiara Imperato, Maria Pagano, and Tiziana Mancini. 2023. "all is fair in… meme!" how heterosexual users perceive and react to memes, news, and posts discriminating against sexual minorities. *Social Sciences*, 12(2):74.

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.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

Kyle Langvardt. 2017. Regulating online content moderation. *Geo. LJ*, 106:1353.

Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2020. Mapping twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, 39(7):711–721.

Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2970–2979.

Cristina Moreno-Almeida. 2021. Memes as snapshots of participation: The role of digital amateur activists in authoritarian regimes. *New Media & Society*, 23(6):1545–1566.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Mustafa Oz, Akan Yanik, and Mikail Batu. 2023. Under the shadow of culture and politics: Understanding lgbtq social media activists' perceptions, concerns, and strategies. *Social Media+ Society*, 9(3):20563051231196554.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

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. *arXiv preprint arXiv:2109.05184*.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

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

Victoria Sherratt. 2022. Towards contextually sensitive analysis of memes: Meme genealogy and knowledge base. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), Vienna, Austria*, pages 5871–5872.

Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. 2023. Parameter-efficient long-tailed recognition. *arXiv preprint arXiv:2309.10019*.

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.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. 2023. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812.

A Appendix

A.1 Topic Modeling

We applied the Sparse Additive Generative Models of Text (SAGE) (Eisenstein et al., 2011) topic modeling technique to identify noteworthy words across various class labels within our dataset. We set the hyperparameters max_vocab_size to 1000 and base_rate_smoothing to 1. Tables 7 and 8 present the most notable words for each class as identified by SAGE along with their corresponding salience scores. Among hate targets, words like 'rowling', 'shapiro', 'conservative', 'gays', 'bethesda', and 'corporations' are assigned high scores by SAGE, helping identify the most targeted entities for each label. Within samples labeled *Support*, words such as 'comfortable', 'expression', and 'supportive' hold relevance as they convey acceptance and support.

A.2 Data Filtering

We screened the images according to the following criteria:

- **Irrelevant Images**: We curated images relevant to the LGBTQ+ movement and discarded non-relevant images.
- **All Text or no Text Images**: We discarded images that did not contain significant visual content or did not have any embedded text.
- **Non-English Text**: We majorly collected images that had English content, however, some images may contain non-English words. Our OCR tool was set to English ensuring that only English text was extracted from images.
- **Low-Quality Images**: We discarded images that were highly distorted, blurred, or degraded. We also removed images with illegible text.

Hate		Target			
No Hate	**Hate**	**Undirected**	**Individual**	**Community**	**Organization**
brooke (0.919)	woke (0.306)	father (1.111)	rowling (2.122)	transphobes (1.002)	bethesda (2.649)
envy (0.917)	conservative (0.286)	oppression (1.053)	shapiro (1.814)	terfs (0.934)	corporations (2.334)
comfortable (0.859)	children (0.279)	event (1.043)	ben (1.754)	conservatives (0.846)	companies (2.322)
expression (0.842)	warning (0.279)	center (0.932)	biden (1.733)	turning (0.759)	disney (2.265)
subscribers (0.820)	marriage (0.204)	bigot (0.920)	walsh (1.595)	gays (0.758)	russia (2.232)

Table 7: Topic Modeling for Hate and Target classification tasks. We report the top 5 words for every label in each task sorted according to their salience score.

Stance			Humor	
Neutral	**Support**	**Oppose**	**Humor**	**No Humor**
envy (1.012)	comfortable (1.073)	warning (0.629)	envy (0.480)	risk (0.826)
min (0.964)	subscribers (1.032)	walsh (0.624)	femboy (0.418)	comfortable (0.723)
content (0.938)	brooke (1.000)	matt (0.615)	miss (0.352)	walsh (0.720)
thinks (0.845)	expression (0.946)	replies (0.601)	mematic (0.342)	youth (0.720)
republican (0.804)	supportive (0.907)	oppressed (0.599)	thinks (0.333)	protect (0.640)

Table 8: Topic Modeling for Stance and Humor classification tasks. We report the top 5 words for every label in each task sorted according to their salience score.

A.3 Implementation Details

We conducted all our experiments on Pytorch 2.1.2 combined with an NVIDIA Tesla T4 GPU with 16 GB of dedicated memory. We set the batch size to 16 and trained each model for 10 epochs while monitoring validation AUROC to save the best model for each run. Under these settings, training and validating MemeCLIP for one epoch takes 12 minutes and occupies 7 GB of dedicated memory.

We empirically found the most optimal learning rate for each model. We used a learning rate of 10^{-5} for ViT-L/14 and CLIP Image-Only. We used a learning rate of 5×10^{-5} for BERT and CLIP Text-Only. We used a learning rate of 10^{-3} for CLIP and CLIP-Adapter. For MOMENTA, HateCLIPper, and ISSUES, we used the default settings set by their respective authors. For MemeCLIP, we set the learning rate to 10^{-4}. We set the scaling factor σ for the cosine classifier to 30, and the residual ratio α for the Feature Adapters to 0.2.

Our framework, MemeCLIP, was built upon the base CLIP model provided by OpenAI's official CLIP library[5]. We also implemented CLIP, CLIP Image-Only, and CLIP Text-Only models by using this library. We implemented the Visual Transformer model by using the timm library provided by Huggingface[6]. We implemented BERT by using the Huggingface Transformers library[7]. We obtained the code released by the authors of CLIP-Adapter[8], MOMENTA[9], HateCLIPper[10], and ISSUES[11] to test their respective methods. The total number of parameters for each method is listed in Table 9.

Method	Number of Parameters (M)
ViT-L/14	307
CLIP Image-Only	307
CLIP Text-Only	123
CLIP ViT-L/14	428
CLIP-Adapter	428
MOMENTA	358
HateCLIPper	430
ISSUES	452
MemeCLIP	430

Table 9: Number of Parameters for each implemented method.

[5]https://github.com/openai/CLIP
[6]https://github.com/huggingface/pytorch-image-models
[7]https://github.com/huggingface/transformers
[8]https://github.com/gaopengcuhk/CLIP-Adapter
[9]https://github.com/LCS2-IIITD/MOMENTA
[10]https://github.com/gokulkarthik/hateclipper
[11]https://github.com/miccunifi/ISSUES