Rethinking Multimodal Content Moderation from an Asymmetric Angle with Mixed-modality

Jialin Yuan [†1], Ye Yu [†2], Gaurav Mittal [2], Matthew Hall [2], Sandra Sajeev [2], Mei Chen [2]

[1]Oregon State University, [2]Microsoft Inc.

{yuanjial}@oregonstate.edu

{yu.ye,gaurav.mittal,mei.chen,mathall,ssajeev}@microsoft.com

Abstract

There is a rapidly growing need for multimodal content moderation (CM) as more and more content on social media is multimodal in nature. Existing unimodal CM systems may fail to catch harmful content that crosses modalities (e.g., memes or videos), which may lead to severe consequences. In this paper, we present a novel CM model, Asymmetric Mixed-Modal Moderation (AM3), to target multimodal and unimodal CM tasks. Specifically, to address the asymmetry in semantics between vision and language, AM3 has a novel asymmetric fusion architecture that is designed to not only fuse the common knowledge in both modalities but also to exploit the unique information in each modality. Unlike previous works that focus on representing the two modalities into a similar feature space while overlooking the intrinsic difference between the information conveyed in multimodality and in unimodality (asymmetry in modalities), we propose a novel cross-modality contrastive loss to learn the unique knowledge that only appears in multimodality. This is critical as some harmful intent may only be conveyed through the intersection of both modalities. With extensive experiments, we show that AM3 outperforms all existing state-of-the-art methods on both multimodal and unimodal CM benchmarks.

1. Introduction

With the proliferation of multimodal social media and online gaming, user-generated content followed by recent AI-generated content (e.g., via DALL-E [40], GPT-3 [5], ChatGPT [49] *etc*.) can spread across the internet at a faster rate than ever. While this enables free speech and facilitates information exchange, it comes with the risk of misuse for fake news [35,51] and hate speech [10,44].

Leaving harmful content on social platforms can lead
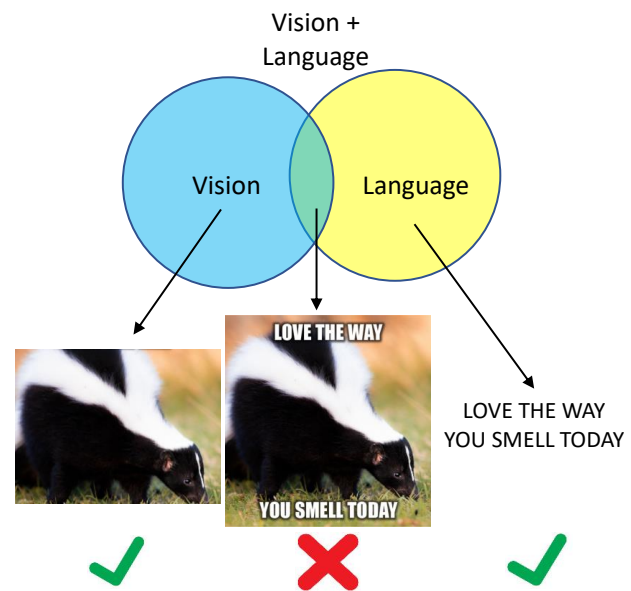
---

† Equal contributions.

Figure 1. An example of a mean meme from Hateful Memes [20] for illustrative purposes. The unimodal vision and language are both benign while the multimodal meme is sarcastic and mean. This is not an actual example of the CM dataset *, which is hateful and would be distasteful to show here.

to harmful consequences, but moderating the tremendous amount of user/AI-generated content on the platforms manually is infeasible due to the large scale and can be harmful to the mental health of human moderators. Therefore, automated content moderation (CM) systems are necessary. There has been extensive research on text-based content moderation [51,54,55]. Recently, there is a study on image-based pornographic content classification and sexual object detection tasks [38]. As social platforms allow the use of different modalities, unsafe multimodal content may evade detection by existing unimodal content moderation systems. Hence, multimodal harmful content detection benchmarks [15,20] have emerged followed by works [9,63] aiming to

automatically detect unsafe multimodal content, including child abuse material, violence, hate speech, sexual content, cyberbullying content, and disinformation [4].

One important form of multimodal content online is memes, which are a combination of image and short text. Understanding memes is a multimodal vision language (VL) task. As noted in previous studies [15], offensive terms by themselves may not necessarily signify hate. It is the overall context that determines whether the intent is harmful or not. Fig. 1 shows an example of mean meme, where the text by itself is just a compliment and the image also seems benign. However, when combining the two modalities the meme becomes sarcastic and mean. This example is for illustrative purpose only. For actual examples which are indeed hateful, please refer to supplementary. To combat the spread of harmful VL content such as hateful memes on social platforms, different VL datasets have been constructed: Facebook proposed a Hateful Memes Challenge and constructed a corresponding dataset [20], which contains memes designed to evade detection by unimodal methods. MMHS150K [15], a large-scale image-text pair dataset originated from Twitter postings, is proposed to benchmark hate speech detection in multimodal publications.

In this work, we approach multimodal (image + text) harmful content detection and propose a novel mixed-modal (a mix of multimodality and unimodality) CM model, Asymmetric Mixed-Modal Moderation (AM3). Image and text are intrinsically different in the information they convey: text is more structured and semantically at a higher level (usually describing the main components of an image while overlooking the subtle details, especially the background). On the other hand, image is unstructured: it is composed of pixels that can provide more low-level details of the context. For example, an image caption is likely to focus on the foreground or the objects of interest in the image. It may contain semantic details like the color or shape of the objects, but unlikely to cover all the details, especially those in the background. We call this *asymmetry in semantics* of VL content. To address this asymmetry, we propose a novel fusion transformer architecture that attempts to maintain the unique knowledge in each modality while fusing the information from the asymmetric semantic levels. As shown in Fig. 1, the knowledge learned from the joint multimodality should contrast that from each unimodality due to this asymmetry in semantics. Sometimes this subtle missing part in unimodality is the determinant for content moderation decisions. We name the discrepancy in the information conveyed by multimodality and each unimodality *asymmetry in modalities*. To tackle this challenge, we propose a novel contrastive loss between the representation learned from multimodality versus each unimodality. In order to learn domain-specific knowledge, we mix multi-

modal dataset with additional unimodal CM datasets in pretraining, similar to [28]. We call this *asymmetry in data* as either modality may be missing in the data, so that the conventional multimodality (each sample contains both modalities) setup becomes mixed-modality (mix of multimodality and unimodality, where each sample may contain both modalities or each unimodality). By including unimodal CM dataset in pretraining, AM3 learns the domain-specific knowledge which helps the model adapt to the downstream tasks. Hence, the downstream CM task performance is improved.

We summarize the main contributions of work below,

- **Asymmetry in semantics**: We propose a novel fusion transformer architecture to fuse different modalities asymmetrically. It enhances the unique knowledge in each modality while effectively fusing the information from the asymmetric semantic levels.

- **Asymmetry in modalities**: We design a novel contrastive loss to squeeze out the distinct knowledge that only exists in multimodality, which is essential in multimodal content moderation.

2. Related Works

Harmful Content Detection. As social media platforms have grown, so have the challenges of content moderation. These challenges have pushed platforms toward automated content moderation as a necessary tool for detecting harmful content. Initially, most of the works are on text [3, 9]. In [10], 25K Tweets are collected and annotated based on whether they contain hate speech keywords or have implicit hate. Logistic regression and SVM are tested to automatically detect hate speech. Besides web crawling data, a large scale machine generated dataset of toxic and benign text statements is provided in [16] using GPT-3 [5]. These labels are then validated by human annotators, and over 95% of the generated toxic labels are legitimately toxic. Over time, images and videos have gained more attention as visual contents are easier to consume and more popular to spread. A large scale dataset for pornographic visual content classification is given in [38]. In [46], videos of conversations are collected as a benchmark for deception detection. Recently, multimodal harmful content detection has attracted more attention. Facebook proposed a Hateful Memes Challenge [20], where each image is associated with a short text. The winner of the challenge [63] outperforms the other competitors significantly by leveraging external labels such as race, age, and entity. Following the same practice, DisMultiHate [26] further improves the performance by disentangling target entities in multimodal memes. Hate-CLIPper [24] proposes a method of intermediate fusion to alleviate the ambiguity alignment between image and text representations. The importance of each
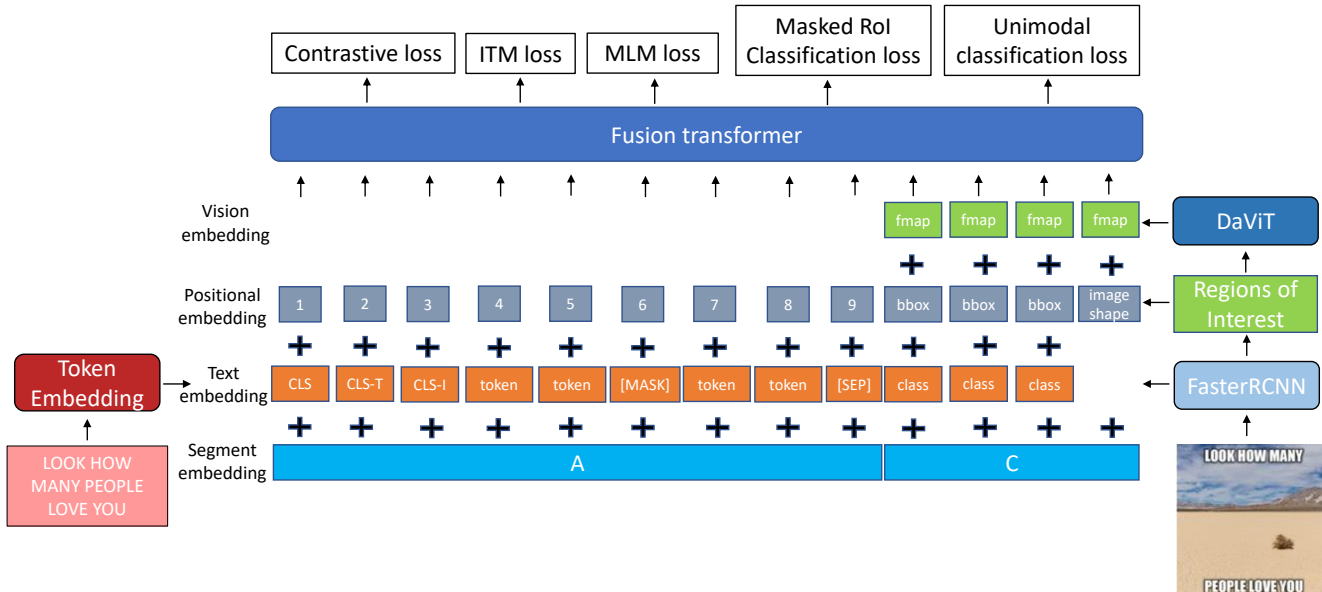
Figure 2. Architecture overview. It shows an example of the pretraining of AM3 with a (T, I) input. For text inputs, we sum up text embeddings, positional embeddings, and the segment embeddings. Visual inputs consist of the text embeddings from detected objects' category labels, the feature map from the vision encoder, positional embeddings, and the segment embeddings. The positional embeddings of visual inputs are computed based on object bounding boxes so that they are permutation invariant to object order.

modality in the Hateful Memes dataset and the robustness of SOTA multimodal classfication algorithms are investigated in [32].

Vision-Language Pretraining. Recent years have witnessed rapid progress in vision-language pretraining (VLP) where vision and language modalities are jointly encoded using a fusion model. The success of BERT [50] inspired many follow-up multimodal fusion models, such as VL-bert [47], VinVL [61], SimVLM [53], and OFA [52], where the text features are concatenated with vision features from image encoder and then fused by BERT or its variants. Besides the masked language modelling (MLM) loss used in BERT pretraining, various loss functions targeting multimodal feature fusion are used, e.g., image-text matching (ITM) loss, region-of-interest (RoI) classification loss. Most of these works learn the joint representation of vision and language through a symmetric feature encoding and fusion process. For example, VL-BERT [47] constructs the multimodal inputs symmetrically where every multimodal feature map has the same components, i.e., text embedding, visual embedding, segment embedding, and positional embedding. Each text embedding is associated to the visual embedding of the entire image while each RoI visual embedding is associated with a dummy text embedding. This simple symmetric architecture enables the fusion of multiple modalities. However, each text token only contains a subset of the entire image. Linking the entire image embedding to it may introduce noise that decreases the performance. On

the other hand, the dummy text embedding does not contain any meaningful information. VinVL [61] simply concatenates text embeddings with the object label embeddings as well as RoI visual embeddings before feeding into the fusion transformer. It assumes that the text embeddings and the visual embeddings share the same (symmetric) level of knowledge and processes them equally.

Recent works on VL foundation models show that dual-encoder architectures can learn strong representation through contrastive objectives on large scale noisy image-text pairs [37, 39, 59]. Florence [59] developed a unified contrastive objective [56] in VLP that enables the model to be adapted for a wide range of vision and VL tasks. Flamingo [2] utilizes an 80B-parameter language model frozen in training and fused with a vision encoder. The huge capacity of Flamingo enables the state-of-the-art performance for few-shot learning.

Our method shares numerous ideas of the previous works mentioned above. However, we pivot to looking at the multimodal content moderation task from an asymmetric angle, both in architecture and data, and target mixed-modality (both multimodal and unimodal) downstream CM tasks. We exploit the discrepancy in vision, language, and multimodal VL pairs, to improve the model capability and training.

3. Method

In this paper, we present a novel fusion transformer architecture pretrained on both VL datasets and unimodality

datasets. To tackle the asymmetry in semantics of CM VL content, we construct vision and language embeddings differently to encourage the model to capture essential knowledge in each modality. Meanwhile, we follow [61] to utilize the object labels from detection as anchors to bridge the language with the corresponding image RoI features. Due to the asymmetry in modality, there is unique knowledge that only exists in the intersection of both modalities. To drive the model to obtain understanding of this, we introduce a novel contrastive loss, *Cross-modality* Contrastive Loss, as part of our pre-training tasks. We use an asymmetric mix of multimodal datasets as well as domain-specific unimodal datasets in pretraining, where a domain-specific classification loss is included to improve downstream task performance.

3.1. Model Architecture for Asymmetry in Semantics

Fig. 2 illustrates the overview architecture of AM3. The model takes mixed modality input: (T, I), (T), or (I), where T represents the text if it exists, and I is the image if it exists. Unlike previous works that try to unify the feature encoding process from both vision and language modalities, we construct the text inputs and visual inputs to the fusion transformer asymmetrically. T is first tokenized through a tokenizer and then fed to a token embedding layer whose outputs are added to positional embeddings and segment embeddings to generate the sequence of linguistic embeddings of text $\mathbf{w}$. The image I is processed as follows: we first use an object detection model to detect objects existing in the image. We also include a bounding box for the entire image (so the bounding box becomes the shape of the image) without an object category associated. For each object, its category label will go through the same token embedding layer as the text input to obtain its text embedding. Its bounding boxes are transferred to the positional embeddings of the RoI through a linear layer. This makes the positional embeddings permutation invariant to the input order of the objects. The visual feature of each RoI is encoded through a feature extractor. We then sum up the text embeddings of object labels, positional embeddings from object bounding boxes, the features from the RoIs, and segment embeddings to obtain the sequence of visual embeddings $\mathbf{v}$. The concatenated pair of (w, v) is fused through a fusion transformer.

3.2. Cross-modality Contrastive Loss for Asymmetry in Modalities

***Cross-modality* Contrastive Loss.** Due to the asymmetry in modalities, the capability of learning the unique knowledge only existing in the intersection of different modalities is critical to content moderation tasks, as demonstrated in Fig. 1. Therefore, we propose the *cross-modality* con-

trastive loss, as given in Equation(1):

$$L_{con} = \max(0, \cos(f_{VL}, f_V)) + \max(0, \cos(f_{VL}, f_L)) \tag{1}$$

where $\cos(\cdot)$ is the cosine similarity function. f_{VL}, f_V, and f_L are the CLS output tokens from the fusion transformer for image + text, image only, and text only, respectively. As shown in Fig. 2, 3 CLS tokens are added to the fusion transformer input. The CLS token is designed to summarize the multimodal knowledge from all tokens while the CLS-I and CLS-T tokens only extract information for vision and language tokens, respectively. By summing up the similarity between f_{VL} vs. f_V, and f_{VL} vs. f_L in the contrastive loss, we push the joint multimodal representation away from the unimodal representations, for the asymmetry in modality, forcing the model to learn the distinct semantic knowledge only in the intersection of both modalities. Fig. 3a illustrates how the attention mask is modified to compute the multimodal and unimodal representations: the attention between CLS and CLS-I/CLS-T as well as between CLS-I and CLS-T tokens are masked out to prevent information leak among different representations. The attention between the CLS-T token and all image tokens are also masked out, and vice versa. In this way, as displayed in Fig. 3b, the CLS output token summarizes the fused information learned from both modalities while CLS-I and CLS-T only contain unimodal information in vision and language, respectively.

Binary classification on domain-related datasets. To help the model effectively adapt to the new domain (CM in our case) when porting a generic model to a specific domain, we include a domain-specific classification loss (L_{domain}) in our pretraining objectives. We collect several content moderation related unimodality datasets discussed in Sec.4.1 into the pretraining corpus. When an input is from these datasets, its CLS output token is projected through a linear layer to predict if the input is harmful or not. We show that this domain-specific classification loss improves downstream performance on CM benchmarks. The domain-specific classification loss is:

$$L_{domain} = -E_{f_{VL}}[\log P(c_d | f_{VL})] \tag{2}$$

where c_d is the domain category label and f_{VL} is the fusion transformer output of the CLS token. In our experiment, we set $c_d = 1$ for harmful inputs while $c_d = 0$ for safe ones. For inputs from generic multimodal VL datasets, we set $c_d = -1$ so that they are ignored in the domain-specific classification task.

As shown in Fig. 2, there are 3 additional pretraining objectives for multimodal fusion: the Masked Language Modeling loss (L_{mlm} on the text tokens similar to [17,22,47,61], the Image-Text Maching loss (L_{itm}) which is computed on the CLS token of joint modalities same as [17, 22], and the Masked RoI classification loss ($L_{roi\text{-}cls}$) similar to [47].
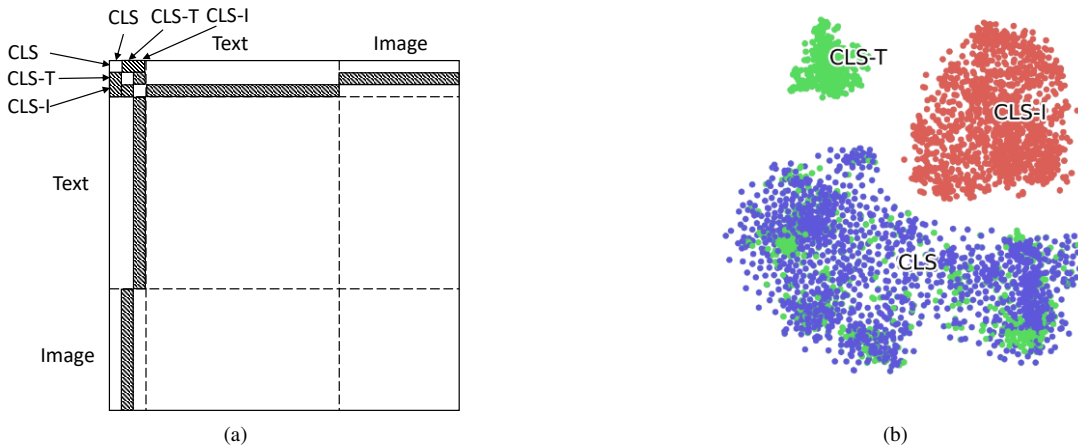
Figure 3. (a): Modified attention mask for contrastive learning. Attention between image tokens (including CLS-I) and CLS-T tokens are masked out, and vice versa. (b): Visualization of the 3 CLS tokens from Hateful Memes after t-SNE [14] reduction.

Overall, our pretraining objective consists of terms as in Equation(3):

$$Loss = \alpha\,L_{con} + \beta\,L_{mlm} + \gamma\,L_{itm} + \lambda\,L_{roi\text{-}cls} + \omega\,L_{domain}$$
(3)

where α, β, γ, λ, and ω are coefficients to balance the various objectives. We set λ to 0.2 and all the other coefficients to 1 throughout the experiments.

4. Experiments

In this section, we first introduce the implementation details. We then discuss the results on downstream CM tasks. Finally, we show an ablation study on the proposed method.

4.1. Implementation Details

Pretraining. As shown in Fig. 2, following [29], we use pretrained FasterRCNN [42] for object detection, but other object detection models, like Yolo [41], can be used as well. We use DaViT [11] as the vision encoder, which encodes the RoIs detected by FasterRCNN into vision embeddings. Both FasterRCNN and DaViT are frozen during training. We use $BERT_{base}$ ($Layers = 12, Hidden\ size = 768, Attention\ heads = 12$) for text embedding and fusion transformer. The model is initialized with pretrained $BERT_{base}$ parameters and optimized using the AdamW optimizer with a base learning rate of 10^{-5} and weight decaying of 10^{-2}. The learning rate was warmed up for 100 training steps and then decayed linearly to zero for the rest of the training. We use a probability of 0.15 in MLM and Masked RoI classification random masking and 0.5 in ITM random replacing. We assign segment tokens 'C' to all visual features. For captions, we set segment tokens to 'A', while for questions and answers, we use 'A' and 'B', respectively. We pretrain the model for $500K$ steps with a batch size of 6144 on 72 NVIDIA V100 GPUs.

Pretraining corpus: We construct our pretraining corpus based on three types of datasets: generic VL multimodal datasets, CM language datasets, and a CM vision dataset.

- **Generic VL multimodal datasets.** We build our corpus from image captioning and visual question-answer datasets, including COCO [30], Conceptual Captions (CC3M) [45], SBU captions [36], Flickr30k [57], CC12M [7], Open-Images [25], GQA [18], and VG-QAs datasets. Following [61], machine-generated captions are used for Open-Images dataset, while captions and question-answer segments are used as text inputs for the other datasets.

- **CM language datasets.** We use 4 language datasets in CM domain: ToxiGen [16], Jigsaw [60], HateXplain [33], and ImplicitHate [12], where we preprocess data so each sample has a harmful or safe label. We use *train* sets in pretraining for ToxiGen, Jigsaw, and HateXplain to avoid data leakage. For text samples without images, we pad [PAD] to vision embeddings. A text classification head predicts the label using the domain-specific classification objective.

- **CM vision dataset.** We use LSPD (Large-Scale Pornographic Dataset) [38] image dataset for CM vision task. Similar to the CM language datasets, we use the *train* set with binary annotation for pretraining and pad [PAD] tokens to the text inputs for fusion transformer. We use an image classification head to predict binary labels using the same classification objective.

Downstream finetuning. All the CM downstream tasks introduced in Sec.4.2 are formulated as classification tasks. The output token on CLS from the fusion transformer is fed into the classification head and trained with cross entropy loss. Hyperparameters including batch size, learning rate,

and training epochs are searched for each task. All classification heads are implemented with an MLP consisting of 2 linear layers and 1 ReLU layer.

Downstream inference. In each task(Sec. 4.2), we utilize the finetuning model and take the classification result from the CLS token as output, the model is named as **AM3**. On the downstream tasks, we conducted 5 experiments with random seeds, reporting their mean and standard variation across the multiple finetuning models. On CM VL tasks, we assessed the model's ability in learning cross-modal knowledge using the best model with unimodal input as **AM3-text** and **AM3-image** respectively. Additionally, we combine the two unimodal results by taking the maximum classification probability as the predicted outcome, and refer to it as **AM3-max**.

4.2. Downstream Datasets

To validate the effectiveness of AM3, we adapt the pretrained model over the content moderation tasks in different modalities.

For CM VL tasks, we adopted Hateful Memes, MMHS150K, and Fakeddit datasets.

- **Hateful Memes [20].** The Hateful Memes dataset consists of more than 10,000 memes, some of which are specially designed so that the text phrases and images are benign when considered separately, but hateful when combined. Therefore, the typical unimodal methods cannot yield good performance on them. To compare with prior works [26, 63], we use 2 different setups: (1) we finetune our model on the *train* set and evaluate on the *dev seen* set. (2) We finetune our model on the combination of *train* and *dev unseen* sets and evaluate on the *test unseen* set. The task uses the Area under Receiver Operating Characteristic curve (AUROC) and accuracy metrics.

- **MMHS150K [15]** The MMHS150K dataset is based on Twitter data consisting of both image and text. We perform binary classification to decide whether a sample is hate or non-hate. We finetune on the *train* and *val* sets and evaluate on the *test* set using F1-score, AUROC, and accuracy metrics.

- **Fakeddit [35].** The Fakeddit dataset is a large-scale multimodal fake news dataset that consists of over 1 million submissions from Reddit, a social news and discussion website where users can post submissions on various subreddits. 2-way, 3-way, and 6-way labels are provided for each sample. We follow the official dataset partition to only use multimodal samples. We focus on the 2-way classification and finetune our model on the *train* set. We compute accuracy on the *val* and *test* sets.

For CM text tasks, we use ToxiGen, HateXplain, and Jigsaw datasets.

- **ToxiGen [16].** ToxiGen is a machine-generated dataset using the massive pretrained language model GPT-3 [5]. The dataset is designed to focus on creating hard-to-classify implicit abusive content in 13 minority groups. We use its *train* and *test* sets. The objective of the task is to predict if each sample is toxic or not and it is evaluated with AUROC.

- **HateXplain [33].** The HateXplain dataset is constructed by collecting posts from Twitter and Gab for research on Explainable Hate Speech Detection. The task is evaluated using AUROC, accuracy, and F1-score.

- **Jigsaw [43].** The Jigsaw dataset is created using comments from Civil Comments for researchers to develop models to recognize toxicity and minimize this type of unintended bias with respect to mentions of identities, including gender, sexual orientation, and religious identity, *etc*. We use the *train* and *test-public* splits for training and testing, respectively. AUROC is computed for evaluation.

We use LSPD for CM vision task.

- **LSPD [38].** LSPD is constructed for visual pornography classification with 5 categories: porn, hentai, drawing, sexy, and non-porn. We followed the porn/non-porn binary classification approach as [38], where the classes 'Hentai' and 'Porn' are grouped as 'porn', while all other classes were labeled 'non-porn' in the binary setting. To evaluate algorithms, accuracy, precision, and recall are measured.

Table 1. Comparisons to the state-of-the-art methods on Hateful Memes.

Method	Dev seen		Test unseen	
	AUROC	Accuracy	AUROC	Accuracy
ERNIE-VIL [58]	78.7	69.0	-	-
Uniter [8]	78.0	68.6	79.1	74.1
VILLA [13]	78.5	71.2	80.0	75.1
VL-BERT [47]	78.8	71.4	79.5	74.5
DisMultiHate [26]	82.8	75.8	-	-
AM3-text(Ours)	59.1	65.2	62.2	64.1
AM3-image(Ours)	44.8	63.0	60.3	63.1
AM3-max(Ours)	56.7	64.3	64.0	64.5
AM3(Ours)	**83.18**(±**0.19**)	**75.98**(±**0.67**)	textbf83.35 (± 0.23)	**76.95**(±**0.36**)

4.3. Result Analysis

Performance comparison on VL tasks: (1) Results on the Hateful Memes comparing to the state-of-the-art approaches are shown in Table 1. We compare to the challenge winner's solutions discussed in [63]: ERNIE-Vil [58], UNITER [8], VILLA [13], and VL-BERT [47], where

Table 2. Comparisons to the state-of-the-art methods on MMHS150K.

Method	AUROC	Accuracy
TKM [15]	73.1	68.2
SCM [15]	73.2	68.5
FCM [15]	73.4	68.4
AM3-text(Ours)	72.7	68.3
AM3-image(Ours)	52.0	52.5
AM3-max(Ours)	72.2	67.7
AM3(Ours)	**74.2 ($\pm$0.09)**	**68.57($\pm$0.79)**

Table 3. Comparisons to the state-of-the-art methods on Fakeddit.

Method	Val acc.	Test acc.
BERT+ResNet50 [35]	89.3	89.1
MVAE+ [27]	-	90.1
MDID [23]	90.8	91.0
EMAF [27]	-	92.3
AM3-text(Ours)	82.23	82.41
AM3-image(Ours)	76.2	75.9
AM3-max(Ours)	83.1	83.3
AM3(Ours)	**93.04($\pm$0.21)**	**93.2($\pm$ 0.11)**

the results are reproduced in [26]. We also compare to the state-of-the-art solution, DisMultiHate [26]. Furthermore, comparing to the baselines: AM3-text, AM3-image, and AM3-max, the AM3 is at least 30% better in terms of AUROC score, which verifies the efficiency of cross-modality understanding in the finetuned model. We adopt the same data augmentation method proposed in [63]: we use Google Vision Web Entity Detection [31] to generate entity tags of each image used as part of the text input. (2) The MMHS150K result is shown in Table 2. We compare to the Feature Concatenation Model (FCM), the Spatial Concatenation Model (SCM), and Texual Kernels Model (TKM) discussed in [15], where they are all CNN + RNN models. It is worth noting that the AM3-text achieves a 72.7 AUROC score, indicating that the dataset predominantly relies on its text modality. Merely considering the maximum classification probability to combine the image modality results leads to a decrease in performance. (3) The Fakeddit result is shown in Table 3. In [35], the authors of the Fakeddit dataset utilize BERT and ResNet50 to encode language and vision, respectively, and then use max-pooling to fuse the multimodal features. MAVE [19] is enhanced with BERT in [27], which is denoted as MAVE+ in the table. EMAF [27] is set up with $BERT_{large\_uncased}$ ($Layers = 24, Hidden\ size = 1024, Attention\ heads = 16$), which is computationally more expensive than our method. The AM3 outperforms unimodal results, AM3-text, AM3-image, and AM3-max,by at least 12.0% in terms of accuracy. On all three datasets, our proposed method achieves the best performance against previous state-of-the-art works. This demonstrates the efficacy of the proposed asymmetric mixed-modal approach. It effectively captures the distinct information that only appears in the intersection of modalities, which is critical in CM decision-making.

Performance comparison on text tasks: AM3 can also handle unimodal CM tasks. We first evaluate our approach on the content moderation text tasks. (1) Results on ToxiGen Classification are listed in Table 4, where we compare to HateBERT [6] and ToxDectRoBERTa [62] on the top-k only version of the dataset. (2) Results on HateXplain dataset are shown in Table 5. Adaptive Length Reduction (AdapLeR) [34] is a method based on BERT while optimizing inference speed. BERT, BERT-HateXplain, BERT-MLM, BERT-RP, and BERT-MRP are different BERT variants discussed in [21]. (3) Results on Jigsaw are shown in Table 6, where we compare to the Toxiciology [1] and Limerobot [1], the top 2 solutions on the leaderboard. On all three CM text datasets, our approach outperforms all the state-of-the-art language models, suggesting the efficacy of the proposed method on mixed-modal (both multimodal and unimodal) downstream CM tasks.

Table 4. Comparisons to the state-of-the-art methods on ToxiGen.

Method	AUROC
ToxDectRoBERTa [16]	85.0
HateBERT [16]	88.0
AM3(Ours)	**91.52 ($\pm$ 0.16)**

Table 5. Comparisons to the state-of-the-art methods on HateXplain.

Method	AUROC	Accuracy	F1
AdaptLeR [34]	-	68.6	-
BERT [48]	85.1	68.9	68.2
BERT-HateXplain [33]	85.1	69.8	68.7
BERT-MLM [21]	85.4	70.0	67.5
BERT-RP [21]	85.3	70.7	69.3
BERT-MRP [21]	86.2	70.4	69.9
AM3(Ours)	**88.25 ($\pm$0.25)**	**81.17($\pm$0.45)**	**80.37($\pm$0.42)**

Table 6. Comparisons to the state-of-the-art methods on Jigsaw.

Method	AUROC
Limerobot [1]	94.7
Toxiciology [1]	94.7
AM3(Ours)	**95.76($\pm$0.27)**

Performance comparison on vision task: Results on the LSPD dataset are presented in Table 8, where we compare them to the outcomes of different methods discussed in [38]. Our approach outperforms previous state-of-the-art methods in terms of accuracy and obtained the highest recall score. Similar to the CM text datasets, this shows the capability of our mixed-modal method on downstream CM vision task, benefiting from a richer representation space with the mixed-modality pretraining.

4.4. Ablation Study

We selected Hateful Memes and MMHS150K for the ablation study of different design choices. To accelerate the analysis, all ablations are performed on a smaller pretraining corpus (Flickr30k, SBU, and COCO), and we pretrain our model for $50K$ iterations.

Table 7. Ablation study of mixed-modality and *cross-modality* contrastive loss.

Text dataset	Vision dataset	*Cross-modality* Contrastive Loss	Hateful Memes		MMHS150K	
			AUROC	Accuracy	AUROC	Accuracy
no	no	no	80.28($\pm$0.18)	74.28($\pm$0.29)	71.91($\pm$0.02)	67.44($\pm$0.05)
no	no	yes	81.18($\pm$1.0)	74.82($\pm$0.51)	72.76($\pm$0.08)	68.21($\pm$0.05)
no	yes	no	80.65($\pm$0.2)	73.82($\pm$0.36)	72.52($\pm$0.05)	68.03($\pm$0.09)
no	yes	yes	81.49($\pm$1.25)	74.98($\pm$0.46)	72.98($\pm$0.13)	68.18($\pm$0.05)
yes	no	no	82.39($\pm$0.08)	76.38($\pm$0.02)	72.79($\pm$0.21)	68.70($\pm$0.09)
yes	no	yes	82.65($\pm$0.2)	75.92($\pm$0.17)	73.36($\pm$0.07)	68.62($\pm$0.09)
yes	yes	no	82.44($\pm$0.23)	75.5($\pm$0.31)	72.85($\pm$0.10)	68.45($\pm$0.07)
yes	yes	yes	**82.94($\pm$0.15)**	**76.45($\pm$0.36)**	**73.96($\pm$0.08)**	**68.73($\pm$0.05)**

Table 8. Comparisons to state-of-the-art methods on LSPD for binary classification.

Method	Accuray	Precision	Recall
Mask-RCNN	86.70	**98.33**	88.00
YOLOv4	92.59	97.03	87.86
SSD	85.32	94.11	85.64
Cascaded Mask RCNN	92.62	95.01	89.95
CNN classifier	87.22	84.86	90.59
AM3(Ours)	**92.86($\pm$0.03)**	92.85 ($\pm$0.15)	**92.73($\pm$0.14)**

Table 9. Ablation study of fusion architecture design.

Architecture	Hateful Memes		MMHS150K	
	AUROC	Accuracy	AUROC	Accuracy
$arch_{VL-Bert}$	80.26($\pm$0.38)	75.72($\pm$0.62)	73.31($\pm$0.15)	68.54($\pm$0.06)
$arch_{vinVL}$	80.53($\pm$0.23)	75.49($\pm$0.17)	73.39($\pm$0.12)	68.4($\pm$0.07)
$arch_{box-position}$	80.74($\pm$0.29)	75.0($\pm$0.09)	73.43($\pm$0.21)	68.4($\pm$0.11)
AM3(Ours)	**82.94($\pm$0.15)**	**76.45($\pm$0.36)**	**73.96($\pm$0.08)**	**68.73($\pm$0.05)**

Model Architecture. To understand the effect of our proposed asymmetric fusion transformer, we create two fusion transformer variants following VL-Bert ($arch_{VL-Bert}$ [47]) and vinVL ($arch_{vinVL}$ [61]), two symmetric fusion designs. Specifically, $arch_{VL-Bert}$ constructs the multimodal embeddings symmetrically so that each text embedding adds to the visual feature of the entire image while each RoI visual embedding adds to a text embedding of a dummy token. $arch_{vinVL}$ creates multimodal embeddings for fusion transformer by simply concatenating the text embeddings from text input and object detection labels, along with visual embeddings. As shown in Table 9, our proposed asymmetric fusion architecture outperforms both symmetric designs, indicating the efficacy of our asymmetric fusion architecture in response to the *asymmetry in semantics*.

Vision Position Embedding from Bounding Box. To validate the effectiveness of using bounding boxes for positional embeddings, we created a variant using the counting index of the tokens for positional embeddings (used in [47, 61]). As shown in Table 9, positional embeddings generated from bounding box captures the ordering information in image (permutation invariant to the input order). Therefore, it achieves a better performance.

***Cross-modality* Contrastive Loss.** As shown in Table 7, the average baseline scores on Hateful Memes and MMHS150K are 80.28% and 71.91%, respectively, measured in terms of AUROC. By adopting the *cross-modality*

contrastive loss, the score is improved by +1.1% on Hateful Memes and +1.2% on MMHS150K. The significant improvements show that our approach to the *asymmetry in modalities* has a strong capability to capture distinct knowledge from the intersection of different modalities.

Pretraining on Unimodal CM Datasets. Table 7 shows the result $w/$ and w/o the unimodal CM datasets in the pretraining corpus. Using the CM text datasets improves the task scores by +2.6% and +1.2% from baseline, respectively. Using the CM image datasets improves the score by +0.5% and 0.8%, respectively. This shows that introducing *asymmetry in data* into the pretraining stage, with the datasets relevant to the domain, is effective and can improve downstream tasks by a significant margin.

Combination of *Cross-modality* Contrastive Loss and Unimodal CM Datasets. As shown in Table 7, utilizing the CM text dataset and the CM vision dataset together leads to further improvement (+0.6% on Hateful Memes and +0.8% on MMHS150K) in comparison to the best score when using either CM text dataset or CM vision dataset. Adding *cross-modality* contrastive loss on top of the unimodal CM text and vision datasets further improve the performance: when enabling all of these components, we achieve the highest average AUROC score of 82.94% for Hateful Memes and 73.96% for MMHS150K. It indicates the efficacy of our proposed method.

5. Conclusion

In this paper, we present a novel mixed-modal CM model, Asymmetric Mixed-Modal Moderation (AM3), for both multimodal and unimodal content moderation. We propose an asymmetric fusion architecture to fuse multimodal knowledge. Furthermore, we design a novel *cross-modality* contrastive loss to learn the distinct knowledge that can only be conveyed when combining both modalities, which is critical for multimodal CM tasks. Besides using multimodal VL datasets, we also include unimodal CM datasets in pretraining, which not only relaxes data constraints but also improves downstream task performance. With extensive experiments, we show AM3 achieves the new state-of-the art on various multimodal and unimodal CM benchmarks.

References

[1] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 3

[3] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 2

[4] Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online, Nov. 2020. Association for Computational Linguistics. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 6

[6] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020. 7

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 6

[9] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*, 2020. 1, 2

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017. 1, 2

[11] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *arXiv preprint arXiv:2204.03645*, 2022. 5

[12] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*, 2021. 5

[13] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 6

[14] Andrej Gisbrecht, Alexander Schulz, and Barbara Hammer. Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing*, 147:71–82, 2015. 5

[15] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478, 2020. 1, 2, 6, 7

[16] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022. 2, 5, 6, 7

[17] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 4

[18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[19] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019. 7

[20] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020. 1, 2, 6

[21] Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. Why is it hate speech? masked rationale prediction for explainable hate speech detection. *arXiv preprint arXiv:2211.00243*, 2022. 7

[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 4

[23] Armin Kirchknopf, Djordje Slijepčević, and Matthias Zeppelzauer. Multimodal detection of information disorder from social media. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE, 2021. 7

[24] Gokul Karthik Kumar and Karthik Nanadakumar. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*, 2022. 2

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 5

[26] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147, 2021. 2, 6, 7

[27] Peiguang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Transactions on Multimedia*, 2021. 7

[28] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2592–2607, Online, Aug. 2021. Association for Computational Linguistics. 2

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 5

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[31] Google LLC. Cloud vision api: Web entity detection. 2021. Accessed: September 20, 2021. 7

[32] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. 3

[33] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021. 5, 6, 7

[34] Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. Adapler: Speeding up inference by adaptive length reduction. *arXiv preprint arXiv:2203.08991*, 2022. 7

[35] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019. 1, 6, 7

[36] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 5

[37] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*, 2021. 3

[38] Dinh Duy Phan, Thanh Thien Nguyen, Quang Huy Nguyen, Hoang Loc Tran, Khac Ngoc Khoi Nguyen, and Duc Lung Vu. Lspd: A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15(1), 2022. 1, 2, 5, 6, 7

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5

[43] Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. Detecting unintended social bias in toxic language datasets. *arXiv preprint arXiv:2210.11762*, 2022. 6

[44] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. 1

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[46] Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777, 2019. 2

[47] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3, 4, 6, 8

[48] Arvind Subramaniam, Aryan Mehra, and Sayani Kundu. Exploring hate speech detection with hatexplain and bert. *arXiv preprint arXiv:2208.04489*, 2022. 7

[49] Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023. 1

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[51] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018. 1

[52] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3

[53] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[54] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. 1

[55] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. 1

[56] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 3

[57] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[58] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021. 6

[59] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[60] Sara Zaheri, Jeff Leath, and David Stroud. Toxic comment classification. *SMU Data Science Review*, 3(1):13, 2020. 5

[61] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3, 4, 5, 8

[62] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. The ineffectiveness of algorithmic debiasing for toxic language detection. In *EACL*, 2021. 7

[63] Ron Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*, 2020. 1, 2, 6, 7