

ContextAware: A Multi-Agent Framework for Detecting Harmful Image-Based Comments on Social Media

Zheng Wei¹, Mingchen Li², Pu Zhang², Xinyu Liu², Huamin Qu^{1*}, Pan Hui^{2†}

¹The Hong Kong University of Science and Technology

²The Hong Kong University of Science and Technology (Guangzhou)

zwei302@connect.ust.hk, {mli736,pzhang012,xliu055}@connect.hkust-gz.edu.cn,
huamin@cse.ust.hk, panhui@ust.hk

Abstract

Detecting hidden stigmatization in social media poses significant challenges due to semantic misalignments between textual and visual modalities, as well as the subtlety of implicit stigmatization. Traditional approaches often fail to capture these complexities in real-world, multimodal content. To address this gap, we introduce *ContextAware*, an agent-based framework that leverages specialized modules to collaboratively process and analyze images, textual context, and social interactions. Our approach begins by clustering image embeddings to identify recurring content, activating high-likes agents for deeper analysis of images receiving substantial user engagement, while comprehensive agents handle lower-engagement images. By integrating case-based learning, textual sentiment, and vision-language models (VLMs), *ContextAware* refines its detection of harmful content. We evaluate *ContextAware* on a self-collected *Douyin* dataset focused on interracial relationships, comprising 871 short videos and 885,502 comments—of which a notable portion are image-based. Experimental results show that *ContextAware* not only outperforms state-of-the-art methods in accuracy and F1 score but also effectively detects implicit stigmatization within the highly contextual environment of social media. Our findings underscore the importance of agent-based architectures and multimodal alignment in capturing nuanced, culturally specific forms of harmful content.

1 Introduction

The detection of hidden stigmatization and discriminatory content on social media platforms has become an increasingly urgent challenge in the digital age [Mossie and Wang, 2020; Cao *et al.*, 2024; Wang and Lee, 2024]. The ubiquity of user-generated content means that harmful sentiments can proliferate rapidly, often veiled in subtle language, imagery, and cultural expressions that elude traditional detection methods.

Implicit stigmatization manifests through nuanced linguistic features, visual cues, metaphors, sarcasm, and insinuations—elements that pose significant challenges to conventional algorithms.

Traditional content detection techniques predominantly rely on explicit keywords, predefined rules, and basic visual features [Yin *et al.*, 2016]. While such approaches are effective for overtly harmful content, they struggle with subtlety and context-dependency. The complexity intensifies when multimodal content—such as memes, image-based comments, or images with layered references—is involved. Not only must each modality be analyzed, but the intricate interplay among textual, visual, and social interaction cues (e.g., the number of likes) requires consideration. Existing methods often emphasize feature extraction and fusion [Cao *et al.*, 2023] but still face semantic gaps and misalignments between heterogeneous modalities [Yang *et al.*, 2022].

Recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs), such as *GPT-4o*, *BLIP-2*, and *CLIP*, have motivated new attempts to exploit zero-shot or few-shot learning for implicit content detection [Hou *et al.*, 2024; Li *et al.*, 2023; Pourpanah *et al.*, 2022; Ji *et al.*, 2024]. These techniques can generate captions or textual representations of images, facilitating cross-modal alignment and reducing semantic gaps. However, the quality of these representations can vary: if captions are too generic or lack detail, key contextual or visual nuances may remain undetected [Gallegos *et al.*, 2024]. Moreover, social media platforms like Douyin often exhibit rapid and culturally specific evolutions in stigmatizing content, further complicating detection [Mossie and Wang, 2020].

In this work, we address the crucial yet underexplored role of contextual information—video titles, hashtags, comment text, and social interaction features—in detecting implicit stigmatizing content. Specifically, we focus on the challenge of identifying harmful image-based comments under short videos. Each video contains a title and hashtags, while each comment includes either text or an image and an associated “like” count. Traditional methods tend to overlook these cross-modal contextual clues, which are critical for understanding the subtle or implied harmful intent of image-based comments.

To this end, we introduce *ContextAware*, an agent-based framework designed to incorporate these social media con-

*Corresponding authors

†Corresponding authors

textual features into the detection pipeline. We evaluate our approach using a self-collected dataset from *Douyin*, focusing on intimate interracial relationships, a topic where implicit stigmatization frequently occurs but is especially challenging to detect with conventional methods. Our dataset contains 871 short videos and 885,502 comments (including image-based ones). Through extensive experiments, we show that *ContextAware* provides more accurate and robust detection of implicit stigmatizing content than existing state-of-the-art models. By leveraging video-level context, comment sentiment, and social cues such as “like” counts, the framework can uncover subtle negative undertones often overlooked by purely keyword- or feature-based methods.

Our key contributions are summarized as follows:

- **Agent-based Context Integration.** We propose *ContextAware*, a multi-agent framework that aggregates textual, visual, and interaction cues (e.g., “likes”) into a cohesive detection pipeline.
- **Case-based Learning for Harmful Content.** We introduce a novel strategy that clusters analysis results of harmful image-based comments to extract generalizable principles of implicit stigmatization.
- **Real-world Dataset.** We collect a Douyin dataset on interracial relationships, showcasing the prevalence of implicit discriminatory imagery and providing a testbed for context-driven detection.
- **Improved Performance.** Experimental results indicate that *ContextAware* significantly outperforms current baselines on both accuracy and F1 scores, demonstrating the efficacy of an agent-driven approach.

The remainder of this paper is organized as follows. Section 2 reviews related work on implicit content detection and multimodal analysis. Section 4 introduces the *Douyin* dataset. Section 5 describes the experimental setup. Section 6 presents the results and comprehensive analyses. Finally, Section 7 concludes the paper and discusses potential directions for future research.

Stakeholders & Multidisciplinary Collaboration Our study is approved by the IRB of HKUST (GZ) (HKUST (GZ)-HSP-2024-0065), and is carried out by an interdisciplinary author team that brings together expertise in computational media, multimodal machine learning, data science, visualization, and AI to ensure technical rigor, social relevance, and ethical compliance.

2 Related Work

The detection of implicit stigmatization and discriminatory content on social media has become increasingly critical in light of the subtle yet pervasive nature of harmful user-generated posts. While many early methods have contributed to addressing explicit hateful or offensive language, the complexity of implicit stigmatization continues to pose substantial challenges—especially when visual, textual, and contextual cues must be considered jointly.

2.1 Implicit Content Detection in Social Media

Traditional approaches to content detection on social platforms rely on explicit keywords, rules, and handcrafted visual features [Zhu *et al.*, 2016]. Although these methods can flag overtly harmful content, they often fail to capture implicit stigmatization, which frequently takes the form of linguistic nuances or hidden visual cues. Sarcasm, metaphors, and insinuations necessitate a more holistic approach that integrates semantic and contextual understanding. Advancements in deep learning for NLP and computer vision have significantly improved the detection of implicit content. For instance, Rao *et al.* [Rao *et al.*, 2014] combined sentiment analysis with topic modeling to unveil subtle negative emotions embedded in social media comments. Multimodal methods that merge textual and visual information have further enhanced detection performance. Ji *et al.* [Ji *et al.*, 2024] employed cross-modal alignment to detect harmful memes, while Farias *et al.* [Farias *et al.*, 2016] and Zhang *et al.* [Zhang *et al.*, 2019] examined linguistic phenomena that convey implicit biases. Radford *et al.* [Radford *et al.*, 2021] highlighted the advantages of simultaneously analyzing text and images, achieving zero-shot transfer capabilities across various tasks. More recently, Meng *et al.* [Meng *et al.*, 2024] introduced the MMLSCU dataset, leveraging multimodal data and chain-of-reasoning strategies to address live streaming content. Despite these strides, numerous challenges persist—particularly in cross-cultural and cross-linguistic contexts. The variability in how different cultures express implicit stigmatization undermines the generalizability of detection models, and the inherent diversity of user-generated content places high demands on nuanced contextual analysis.

2.2 Zero-Shot Learning and Multi-Agent Systems in Multimodal Analysis

LLMs demonstrate strong zero-shot capabilities, performing diverse tasks without task-specific training data [Hou *et al.*, 2024]. Models such as *GPT-3.5* and *GPT-4o* excel at text generation, sentiment analysis, and question-answering, aided by Reinforcement Learning from Human Feedback (RLHF) [Bai *et al.*, 2022] for more human-like outputs [Singgalen, 2024; Achiam *et al.*, 2023; Li *et al.*, 2024]. In vision-language modeling, parameter scaling has bolstered zero-shot and few-shot performance in question-answering, with models like VisualBERT [Li *et al.*, 2019] and BLIP-2 [Li *et al.*, 2023] attaining state-of-the-art results. Additional techniques—such as PromptHate [Cao *et al.*, 2023] for hate-speech detection and PromptCap [Hu *et al.*, 2022] for enhanced image descriptions—further extend capabilities in domain-specific tasks. However, LLM-based systems often struggle to detect subtle and context-dependent stigmatization, especially when cultural or platform-specific cues significantly shape the user’s intent [Gallegos *et al.*, 2024]. For example, Douyin posts featuring interracial relationships may draw on culturally specific stigmatizing tropes [Wei *et al.*, 2024a], which require precise interpretation of video titles, hashtags, user comments, and the nature of image-based content.

Multi-agent systems have emerged as a promising direction for handling these intricacies. By decomposing the detection

pipeline into specialized agents—each optimized for a particular modality (e.g., text or images) or context (e.g., comment sentiment or video metadata)—it becomes feasible to fuse the outputs of multiple experts into a more comprehensive understanding [Cheng *et al.*, 2024; Hong *et al.*, 2023; Talebirad and Nadiri, 2023; Wu *et al.*, 2023]. Integrating LLMs and VLMs within these systems can further enhance fine-grained analysis: language-processing agents may excel at parsing subtle textual insinuations, while image-focused agents leverage CLIP-like embeddings to identify hidden visual cues. Nevertheless, effective coordination among these agents—and the incorporation of rich social media metadata, such as “like” counts and hashtags—remains a key challenge. Existing work often overlooks these user interaction features and the broader video context. To address these limitations, our approach goes beyond conventional cross-modal pipelines by leveraging a purely LLM-driven solution. Specifically, we systematically integrate comment sentiment, social interaction features, and video-level data to enable more accurate identification of implicit stigmatizing content within complex social media environments. By aligning text and image embeddings and incorporating social signals, our method fills a critical gap in the literature, targeting subtle yet pivotal aspects of implicit stigmatization detection.

3 Datasets

Our research uses data collected from *Douyin* (TikTok in China Mainland), China’s leading short-video social media platform with more than 1 billion registered users and more than 700 million daily active users. The data collection period spanned from November 1 to November 4, 2024, with daily collection intervals of 20 hours. The cleaned data set comprises 871 short videos and their associated 885,502 comments (including image-based comments) posted between September 2018 and November 2024, specifically focusing on interracial intimate relationships. We selected this data set because previous research indicates significant instances of racial and gender discrimination in similar social media data [Wei *et al.*, 2024a; Zhou *et al.*, 2024]. Studies on interracial intimate relationships on Chinese social media reveal that relationships between black men and Chinese women often attract intense negative sentiment, while those involving white women and Chinese men receive comparatively fewer negative comments [Wei *et al.*, 2024a]. These varying attitudes underscore pronounced racial and gender biases, presenting complex and nuanced examples of implicit stigmatization, which makes this data set particularly suitable for evaluating and demonstrating the effectiveness of our proposed *ContextAware* framework. The data set examines two primary categories of intimate interracial relationships: 1. Black women and Chinese men, 2. Black men and Chinese women. To construct the initial corpus, we extracted video data from *Douyin* using the Chinese search terms “黑人老公 (Black Husband)” and “黑人老婆 (Black Wife)”. We used an n-gram model to extract relevant phrases and calculated their cosine similarity with the primary keywords using embedding models. Following similarity-based ranking, we manually selected the top 17 most relevant phrases as keywords for fur-

ther video collection. These keywords included terms such as “非洲老公 (African Husband)”, “中非情侣 (Chinese-African Partners)”, and “喜欢黑人 (Like Black People)”.

After retrieving more than 8,000 short videos through keyword searches, three researchers conducted manual reviews to ensure topical relevance to interracial relationships, ultimately retaining only 871 videos and excluding a large number of others that were irrelevant or duplicated. For the extraction of text- and image-based comments, we developed a Web crawler based on the open-source *MediaCrawler*¹ framework. During data cleaning, we removed username mentions (@+username) while preserving emoji categories to capture emotional expressions within the comments.

Subsequently, two researchers utilized the method of manual labeling combined with embedding for deduplication to mark the image datasets in the video comment sections of both black men and Chinese women and Chinese men and Black women as either harmful or not harmful. The two researchers completed the labeling independently. In case of disputed images, they were removed. Eventually, a total of 6,977 images were obtained from the video comment section of black men and Chinese women (of which 3,046 were marked as harmful and 3,931 as not harmful); and a total of 665 images were obtained from the video comment section of Chinese men and Black women (among which 332 were marked as harmful and 333 as not harmful), as shown in Table 1. A random anonymized data sample 10% and detailed technical specifications are available on GitHub².

Dataset	Harmful	Not harmful	Total
Black men and Chinese women	3046	3931	6977
Chinese men and Black women	332	333	665

Table 1: Statistics of the datasets used

4 Method

4.1 Task Definition

On social media platforms, comments under video posts can be multimodal (textual or image-based), and each comment carries an associated “like” count. These comments, together with the video’s title and hashtags, form a comprehensive context. Existing harmful content detection methods frequently ignore such contextual information, leaving them ill-equipped to detect subtle cues of implicit stigmatization or discrimination, especially in image-based comments. Moreover, cultural nuances can influence how harmful content is encoded—whether through localized memes, symbols, or coded language—making detection considerably more complex. So, we focus on identifying harmful image-based comments by leveraging both the video context (title, hashtags) and the broader discussion environment (textual comments, social interaction indicators). Concretely, let a video post on a social media platform be represented as

¹<https://github.com/NanmiCoder/MediaCrawler>

²<https://github.com/relateddata/ContextAwareDatasetRelease>

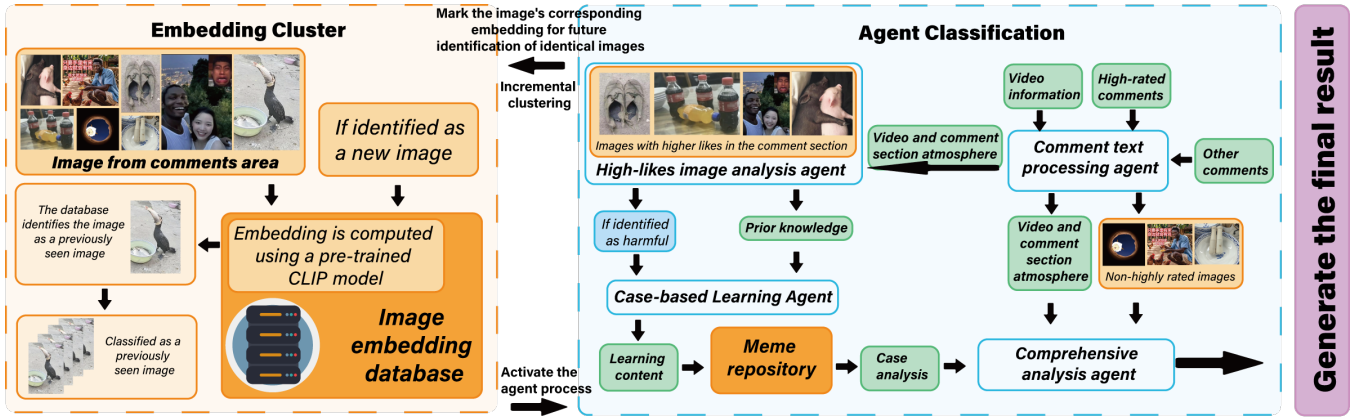


Figure 1: For all images in the comment section, the process begins with embedding clustering. If an image is identified as new, the *ContextAware* system is activated, initiating the agent classification process. During this phase, the highly rated images—those in the top 10% by like count in the comment section of each video (rounded up)—are analyzed, and the results are stored in a meme repository. Finally, non-highly rated images are processed by integrating case studies from the meme repository, comment text analysis, and the image itself, which are passed to the comprehensive analysis agent within *ContextAware* for the final determination of whether the image is harmful.

$V = (T_v, H_v)$, where T_v represents the video title and H_v denotes the video hashtags. Let its associated comment set be, $C = \{c_1, c_2, \dots, c_n\}$, where each comment $c_i = (m_i, s_i)$ has content m_i (text or image) and a social interaction feature s_i (e.g., like count). The task is to learn a function, $f : (V, C) \rightarrow \{0, 1\}$, that determines whether any image-based comment $c_j \in C$ (with m_j containing an image) is harmful (1 for harmful, 0 for non-harmful).

In this work, “harmful” content encompasses messages of hate, stigmatization, or discrimination, which may be explicitly expressed or implicitly encoded in visual symbols and memes. We emphasize *image-based* comments because they can conceal subtle cues (e.g., modified symbols or memes) that cannot be captured by text-only analyses. In addition, we harness the contextual data—such as titles, hashtags, comment text, and “like” counts—to better uncover embedded cultural or topic-specific references that can amplify the harmfulness of certain images, as shown in Figure 2.

4.2 ContextAware Framework

The *ContextAware* framework (Figure 1) is a multi-agent system designed to identify harmful images in video comment sections by incorporating a variety of contextual signals from social media. By harnessing video metadata (titles, hashtags), textual sentiment and semantic cues, and social features (e.g., comment “likes”), our framework provides a more holistic analysis than conventional methods that rely on text-only or image-only indicators. Moreover, it optimizes computational resources through a modular agent design, ensuring redundant tasks (e.g., repeated embeddings of near-duplicate images) are minimized. The framework comprises five core components: (1) *Image Embedding Database*; (2) *Comment Text Processing Agent*; (3) *High-likes Image Analysis Agent*; (4) *Case-based Learning Agent*; (5) *Comprehensive Analysis Agent*. Figure 1 illustrates how the agents interact: after the Comment Text Processing Agent and the Image Embedding Database process the text and images, respectively, intermediate findings are forwarded to the High-likes Image Analysis

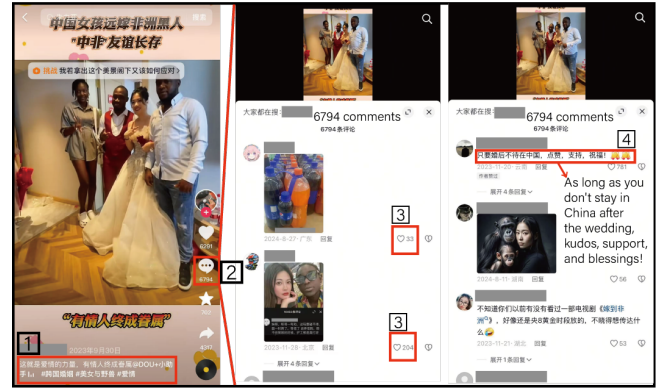


Figure 2: Douyin short video interaction data annotation example: (1) Video title and hashtags (‘Chinese girl marries black African man’, ‘China-Africa friendship endures’). (2) Number of comments on the video (6794 comments). (3) Number of likes on specific image (e.g., 33 and 204). (4) Example of comment text (‘As long as you don’t stay in China after the wedding, kudos, support and blessings!’).

Agent. Images flagged as potentially harmful are then passed to the Case-based Learning Agent, which refines detection principles. Finally, the Comprehensive Analysis Agent consolidates all insights for the final decision.

Image Embedding Database

Before any deep analysis, we employ the *CLIP* model (ViT-B/32) to obtain embeddings for each image, and cluster them with *DBSCAN* to identify near-duplicates (e.g., images differing only by slight cropping or compression artifacts). We set $\text{eps}=0.18$, $\text{min_samples}=2$, and $\text{metric}=\text{cosine}$. These parameters were selected based on a small hyperparameter grid search, balancing the detection of near-identical images against mistakenly clustering distinct images. Near-duplicates are thus handled collectively to reduce redundant computations in subsequent stages. In borderline

cases (e.g., minor stickers, text overlays, or rotations), we lean on *CLIP*'s robust visual-semantic embeddings to maintain grouping consistency.

Comment Text Processing Agent

This agent focuses on deriving a high-level understanding of the comment section's primary topics and sentiment, which serve as auxiliary signals for image interpretation. Specifically, we select the 20 comments with the most "likes," as these typically reflect the dominant themes, and then randomly sample 5 additional comments to capture a more diverse range of viewpoints. We settled on the "20 + 5" strategy after experimenting with smaller samples—such as "10 + 5" and "15 + 5"—on a subset of about ten videos. Although these smaller samples could still identify the overarching sentiment, they occasionally missed crucial minority opinions and nuanced negative sentiment, leading to less accurate or incomplete thematic coverage. By contrast, "20 + 5" consistently captured both mainstream and niche perspectives, improving the agent's understanding of the comment environment without incurring a prohibitive computational overhead. Next, we incorporate the video title and hashtags (T_v, H_v) to contextualize the discussions. Using *GPT-4o* with a zero-shot prompting strategy (temperature=0), the agent infers common topics and sentiment (positive, negative, or neutral), as well as any potential sarcasm or harmful undertones. This textual context helps interpret visually encoded content, as certain cultural references or stigmatizing remarks may only be discernible through a combined reading of text, hashtags, and concurrent social discussions.

High-likes Image Analysis Agent

In the next stage, the High-likes Image Analysis Agent extracts the top 10% most "liked" images under each video. We opted for the top 10% threshold after observing, in a pilot study, that harmful or controversial images often accumulate disproportionately high "like" counts, presumably due to virality or user engagement. This agent integrates the overall sentiment derived from the Comment Text Processing Agent, as well as metadata from $V = (T_v, H_v)$. For each image, the agent performs a preliminary assessment using *GPT-4o* (temperature=0) with a prompt that includes: (i) the textual comments' predominant sentiment, (ii) the video context (title + hashtags), and (iii) any known harmful-content markers (e.g., violent, extremist symbols). This early filtering step narrows down which images need deeper investigation, thereby optimizing computational efficiency.

Case-based Learning Agent

The Case-based Learning Agent performs a more in-depth analysis on images initially flagged as "potentially harmful" by the High-likes Image Analysis Agent. It specifically aims to detect *implicit* stigmatization, such as subtle visual metaphors, coded language overlaid on images, or racial caricatures. We leverage: (1) Semantic & Sentiment Context: From the Comment Text Processing Agent. (2) A curated repository of common stigmatization techniques related to the video topic. For instance, if (H_v) indicates "interracial relationships," we include references to typical racist caricatures or memes. Within this agent, *GPT-4o* (temperature=0.5)

performs nuanced image-text interpretation. It checks for alignment with known stigmatization methods, then logs an analysis result (short natural-language explanation). To organize these results, we employ *SentenceTransformer* (*m3e-base*) to embed each explanation and cluster them using *DBSCAN* (eps=0.5, min_samples=5). Each cluster typically reflects a distinct "theme" of stigmatization (e.g., racial stereotypes, hateful symbology). Finally, these clusters are summarized into a concise set of *principles for identifying harmful images*—for instance, "Caricatured facial features + negative textual sentiment about a specific ethnicity." This distilled knowledge base can then generalize to new images in subsequent processing.

Comprehensive Analysis Agent

The Comprehensive Analysis Agent fuses outputs from all prior agents. It takes: 1) The textual sentiment and semantic context from the Comment Text Processing Agent, The set of harmfulness principles learned by the Case-based Learning Agent, and The original images (now potentially deduplicated by the Image Embedding Database). Using a final *GPT-4o* classification prompt, it weighs textual sentiment, user-engagement signals, the consolidated stigmatization criteria, and the image embeddings to make a definitive decision: harmful or not harmful. In cases where the textual analysis conflicts with the image-based evidence (e.g., negative sentiment but unclear visuals), the system uses the learned principles to resolve ambiguity. This multi-agent synergy—each agent specializing in a distinct facet—allows *ContextAware* to achieve more robust, context-sensitive judgments.

5 Experiments

5.1 Experimental setup

We implement *ContextAware* by employing *GPT-4o* for all LLM-based agents. The Comment Text Processing Agent, High-likes Image Analysis Agent, and Comprehensive Analysis Agent each use a temperature of 0, facilitating deterministic, reproducible outputs. By contrast, the Case-based Learning Agent uses a temperature of 0.5 to induce more exploratory reasoning during implicit stigmatization analysis, a design choice validated through small-scale experiments. All other *GPT-4o* parameters remain at their defaults. We set the random seed to 42 for comment sampling. For the Image Embedding Database, we adopt *CLIP ViT-B/32*. *DBSCAN* is configured with eps=0.18, min_samples=2, and metric= cosine. Similarly, within the Case-based Learning Agent, we use the *SentenceTransformer* model (*m3e-base*) and run *DBSCAN* with eps=0.5, min_samples=5, and metric= cosine. These parameters were set based on grid searches that minimized misclustering of either near-identical images or textual analyses. We conducted all *ContextAware* experiments on a server equipped with an *NVIDIA® GeForce RTX 3080 Ti* GPU, using PyTorch 1.11.0 and CUDA 11.3. For both the *m3e-base* and *CLIP ViT-B/32* models, we employed the official Hugging Face configurations.

Dataset Composition

For our experiments, we use a self-collected dataset of 871 short Douyin videos focusing on interracial relationships,

accompanied by 885,502 comments (including image-based ones). Harmful vs. non-harmful labels were assigned by human annotators following internal guidelines.

5.2 Baselines

We compared our method with several SOTA multimodal approaches. These multimodal methods include *Concat-BERT*, *MMBT* [Kiela *et al.*, 2019], *MO-MENTA* [Pramanick *et al.*, 2021], *CLIP ViT-L/14* [Dong *et al.*, 2022], and *BLIP-2 Flan-T5-xxl* [Yang *et al.*, 2024]. These methods employ various fusion strategies and model architectures to comprehensively evaluate the performance of our approach in multimodal tasks.

6 Results and Analysis

6.1 Comparison with Baselines

Table 2 compares the overall performance of our proposed approach with SOTA techniques across both datasets. In the evaluation process, the *ContextAware* model performed exceptionally well in handling multimodal data from both Chinese and Black male and female populations. For the Chinese male and female dataset, *ContextAware* ACC of 0.787, an F1 score of 0.676, and a MMAE of 0.213, which were significantly higher than all other models. In the Black male and female dataset, *ContextAware* showed consistent outstanding performance with an accuracy of 0.787, an F1 score of 0.676, and MMAE of 0.213, demonstrating its ability to maintain a good balance between precision and recall, while also exhibiting the lowest average error.

In contrast, the *BLIP-Flan-T5-xxl* model performed reasonably well in terms of accuracy, especially on the Chinese male and female dataset with an accuracy of 0.619. However, its F1 score was relatively low, with values of 0.270 and 0.193, indicating a deficiency in balancing precision and recall, which affected its overall performance. Moreover, the *BLIP-Flan-T5-xxl* model had a higher MMAE, with 0.500 for both the Black male and female dataset and the Chinese male and female dataset, further highlighting its limitations in this task. The *ConcatBERT* model performed poorly across all evaluation metrics. Its accuracy on the Chinese male and female dataset was 0.319, and on the Black male and female dataset, it was 0.499, both the lowest among all models. Its F1 score and MMAE were also lower than other models, indicating that *ConcatBERT* struggled to capture key information in the data and performed inadequately for this task.

In conclusion, the *ContextAware* model clearly outperformed all other models across all evaluation metrics, demonstrating superior performance in terms of accuracy, F1 score, and MMAE. This makes *ContextAware* the best choice for handling multimodal datasets, as it can better understand and process complex language and visual information. In contrast, the *BLIP-Flan-T5-xxl* and *ConcatBERT* models showed weaker performance, particularly in balancing precision and recall and minimizing error, suggesting room for improvement.

6.2 Ablation analysis

Through the ablation study (Table 3), we found that each component contributes significantly to the overall perfor-

Model	Black men and Chinese women			Chinese men and Black women		
	ACC	F1	MMAE	ACC	F1	MMAE
ConcatBERT	0.319	0.361	0.555	0.499	0.348	0.496
MO-MENTA	0.547	0.354	0.333	0.499	0.333	0.500
MMBT	0.564	0.361	0.436	0.501	0.334	0.499
CLIP						
ViT-L/14	0.479	0.491	0.513	0.442	0.493	0.527
BLIP-2						
Flan-T5-xxl	0.619	0.270	0.500	0.535	0.193	0.500
ContextAware	0.909	0.785	0.091	0.886	0.845	0.114

Table 2: Results Overview: Proposed Methodology vs. Baseline Approaches

mance on both datasets. For example, removing the video context and summarized discriminatory principles from *ContextAware* led to a notable decrease in the F1 scores for both tasks. Therefore, we conclude that the effectiveness of *ContextAware* stems from the integration of all its modules, as they collectively enhance performance.

Method	Black men and Chinese women			Chinese men and Black women		
	ACC	F1	MMAE	ACC	F1	MMAE
LLM without agent	0.294	0.284	0.706	0.359	0.358	0.641
ContextAware -no Case based Learning Agent	0.772	0.667	0.228	0.609	0.493	0.391
ContextAware -ALL	0.909	0.785	0.091	0.886	0.845	0.114

Table 3: Ablation analysis on both tasks

6.3 Case study and error analysis

In the successful example on the left side of Figure 3, the system leverages multi-agent collaboration and multimodal alignment to accurately identify the cross-racial marriage metaphor implied by the juxtaposition of the dark beverage bottle and the yellow liquid bottle. By analyzing the comments, it captures high-frequency emoji such as “blackface” and “bomb,” which reveal sarcastic and racially biased sentiments, correctly classifying the content as harmful and demonstrating its capability to detect “metaphorical symbols” and “cultural and societal impacts.” However, in the misclassification example on the right side of Figure 3, the system exposes shortcomings in handling complex contextual classification: the image embedding module fails to recognize the Chinese slang connotation of “worn-out shoes”, which implies unfaithfulness or improper relationships, thus overlooking the humiliating expression toward cross-cultural marriage. Overall, while the successful example highlights the potential of multimodal collaboration and metaphor recognition, the failed case underscores the need to strengthen fine-grained visual feature extraction, refine cross-modal contextual consistency, and enhance the detection of metaphors and indirect discriminatory language to further improve the robustness and generalization of multimodal harmful content detection.



Correct Prediction			Incorrect Prediction		
<div>Meme</div> <div></div>			<div>Meme</div> <div></div>		
GT:Harmful			Pred:Harmful		
Video Title: # Black husbands			Video Title: #MyTikTokLifeDiary #CrossBorder Marriage #AfricanInternational Student A Happy Sino-African Cross-Border Marriage"		
Analysis	This image presents a symbolic and suggestive arrangement: a dark-colored Coca-Cola bottle placed closely next to a bottle containing a yellow liquid. Within the context of the video and its accompanying comments, this combination may be interpreted as an allusion to the interracial marriage relationship depicted, indirectly referencing Black individuals and Chinese individuals through the colors of the beverages (dark for Coca-Cola and yellow for the other drink). This aligns with the direction of racial metaphor. Particularly in a comment section rife with negativity and subtle racial discrimination, such an image is more likely to be exploited as a tool for sarcasm or division, insinuating racial and cultural dynamics. Therefore, based on the given principles, the image adheres to the criteria of "symbolic elements and metaphorical significance" as well as "cultural and social impact," and is deemed harmful—a depiction that belittles and potentially attacks interracial relationships.				
	The video titled "#BlackHusband" suggests that the theme revolves around interracial marriage, particularly the relationship between Chinese individuals and Black individuals. Judging from the comments, the overall atmosphere in the comment section leans negative, with many comments employing sarcasm, irony, and negative emoticons (such as "black face," "bomb," etc.) to express dissatisfaction or a lack of understanding toward such interracial relationships. Some comments directly mock and express skepticism about the viability of these relationships, suggesting they are unlikely to last (e.g., "You'll cry later," "You'll regret it when you get home"). Additionally, there are comments with racially discriminatory undertones (e.g., "Please don't get divorced, so you don't spread viruses to domestic men. Thank you, thank you, thank you"), indicating that negative sentiments or prejudices against Black individuals still persist in this region. Although there are a few neutral comments emphasizing that personal choice should be respected (e.g., "It's her own business who she marries"), such voices are relatively scarce.				
Video Summary	The image depicts a pair of worn-out shoes with cobwebs, and there are no apparent signs of harmful intent or offensive symbols directly related to racial or cultural themes. The visual content does not contain exaggerated or distorted features targeting any specific group or culture. In terms of contextual meaning, such an image could metaphorically imply old-fashioned ideas or neglect, but it lacks explicit elements that align with the principles of identifying harmful imagery. Therefore, based on the provided principles for determining harmful images, including language, visual expression, and symbols, this image does not meet the criteria for being classified as harmful. While the negative sentiment in the video's comment section suggests potential biases among viewers, the image itself does not directly contribute to discrimination or hostility and is more appropriately classified as non_harmful.				
	The video title and thumbnail present a story about a cross-cultural marriage between a Chinese person and an African individual, seemingly portraying a positive depiction of their happy life together. However, the overall tone in the comments section leans more toward negativity. While some highly upvoted comments express support, such as encouragement for "freedom in love," the majority of the comments carry a tone of sarcasm, discrimination, and negativity. For instance, comments like "expanding their social circle" and "support, better than marrying those penniless Chinese men" reveal underlying biases and ironic expressions. Terms such as "charcoal" reflect racially discriminatory undertones. Additionally, remarks like "just don't post this" and "China is not your colony" demonstrate strong xenophobic sentiments and dissatisfaction with cross-cultural marriages. Overall, the comments section is dominated by negative and discriminatory voices.				

Figure 3: The meme predictions are divided into “Correct Prediction” and “Incorrect Prediction” groups, each containing columns like Meme, GT (Ground Truth), Pred (Prediction), Video Title, Analysis, and Video Summary, where the former group shows accurate classifications and the latter highlights misclassifications.

7 Conclusion

In this paper, we introduced and evaluated *ContextAware*, a multi-agent framework for detecting harmful image-based comments within the highly contextual environment of social media video posts. Our results demonstrate that *ContextAware* significantly outperforms SOTA baselines across multiple metrics on two challenging datasets involving interracial relationships. Ablation studies further confirm that each component—particularly the Case-based Learning Agent and the cross-modal principles it summarizes—contributes notably to the overall performance, underscoring the importance of combining contextual cues (e.g., video titles, hashtags, and user comments) with advanced multimodal alignment strategies.

ETHICS AND BOARDER IMPACTS

Our study is fully reproducible: Sections 5 and 6 detail the experimental framework, hyper-parameter settings, and overall setup, and all source code is openly available on GitHub³. The dataset is drawn from publicly accessible online platforms and contains no personally identifiable information; all samples are anonymized, and only random subsets are

³<https://github.com/shuiyuetingdong/ContextAware>

shared. Our sole aim is to detect potentially stigmatizing imagery in discussions of interracial relationships, not to reinforce stereotypes, and we have taken multiple steps to minimise inadvertent bias. The project was approved by the Institutional Review Board under protocol [HKUST(GZ)-HSP-2024-0065]. Looking ahead, we anticipate that AI systems, including large language models, will play an expanding role in moderating social-media content [Liao *et al.*, 2025; Wei *et al.*, 2024b], raising debates about standards for harmful-content detection and the ethics of delegating such judgments to machines [Wang *et al.*, 2024; Wei *et al.*, 2024b; Cheong *et al.*, 2024]. Continued research and the inclusion of human oversight remain essential for ensuring safety and reliability when AI is used to identify harmful or stigmatizing media [Schwartz *et al.*, 2022].

Acknowledgments

This work is supported by the Guangdong Provincial Talent Program, Grant No.2023JC10X009.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni

- Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [Bai et al., 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [Cao et al., 2023] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*, 2023.
- [Cao et al., 2024] Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4575–4584, 2024.
- [Cheng et al., 2024] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*, 2024.
- [Cheong et al., 2024] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding human values: rethinking us law for generative ai’s societal impacts. *AI and Ethics*, pages 1–27, 2024.
- [Dong et al., 2022] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*, 2022.
- [Farías et al., 2016] Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):1–24, 2016.
- [Gallegos et al., 2024] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.
- [Hong et al., 2023] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- [Hou et al., 2024] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer, 2024.
- [Hu et al., 2022] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.
- [Ji et al., 2024] Junhui Ji, Xuanrui Lin, and Usman Naseem. Capalign: Improving cross modal alignment via informative captioning for harmful meme detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4585–4594, 2024.
- [Kiela et al., 2019] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.
- [Li et al., 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [Li et al., 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [Li et al., 2024] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. “hot” chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36, 2024.
- [Liao et al., 2025] Junxiang Liao, Zheng Wei, Zeyu Yang, Xian Xu, Pan Hui, Changyang He, and Muzhi Zhou. “even when success seems impossible, i keep streaming”: How do chinese elderly streamers interact with platform algorithmic (in) visibility. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2025.
- [Meng et al., 2024] Zixiang Meng, Qiang Gao, Di Guo, Yunlong Li, Bobo Li, Hao Fei, Shengqiong Wu, Fei Li, Chong Teng, and Donghong Ji. Mmlscu: A dataset for multi-modal multi-domain live streaming comment understanding. In *Proceedings of the ACM on Web Conference 2024*, pages 4395–4406, 2024.
- [Mossie and Wang, 2020] Zewdie Mossie and Jenq-Haur Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087, 2020.
- [Pourpanah et al., 2022] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.
- [Pramanick et al., 2021] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multi-modal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*, 2021.

- [Radford *et al.*, 2021] 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.
- [Rao *et al.*, 2014] Yanghui Rao, Qing Li, Xudong Mao, and Liu Wenxin. Sentiment topic models for social emotion mining. *Information Sciences*, 266:90–100, 2014.
- [Schwartz *et al.*, 2022] Reva Schwartz, Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. *Towards a standard for identifying and managing bias in artificial intelligence*, volume 3. US Department of Commerce, National Institute of Standards and Technology, 2022.
- [Singgalen, 2024] Yerik Afrianto Singgalen. Analyzing an interest in gpt 4o through sentiment analysis using crispdm. *Journal of Information Systems and Informatics*, 6(2):882–898, 2024.
- [Talebirad and Nadiri, 2023] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.
- [Wang and Lee, 2024] Han Wang and Roy Ka-Wei Lee. Memecraft: Contextual and stance-driven multimodal meme generation. In *Proceedings of the ACM on Web Conference 2024*, pages 4642–4652, 2024.
- [Wang *et al.*, 2024] Jingwei Wang, Ziyue Zhu, Chunxiao Liu, Rong Li, and Xin Wu. Llm-enhanced multimodal detection of fake news. *PloS one*, 19(10):e0312240, 2024.
- [Wei *et al.*, 2024a] Zheng Wei, Yixuan Xie, Danyun Xiao, Simin Zhang, Pan Hui, and Muzhi Zhou. Social media discourses on interracial intimacy: Tracking racism and sexism through chinese geo-located social media data. In *Proceedings of the ACM on Web Conference 2024*, pages 2337–2346, 2024.
- [Wei *et al.*, 2024b] Zheng Wei, Xian Xu, and Pan Hui. Digital democracy at crossroads: A meta-analysis of web and ai influence on global elections. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1126–1129, 2024.
- [Wu *et al.*, 2023] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.
- [Yang *et al.*, 2022] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3492–3500, 2022.
- [Yang *et al.*, 2024] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. Mm-instructeval: Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. *arXiv preprint arXiv:2405.07229*, 2024.
- [Yin *et al.*, 2016] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu. Text detection, tracking and recognition in video: a comprehensive survey. *IEEE Transactions on Image Processing*, 25(6):2752–2773, 2016.
- [Zhang *et al.*, 2019] Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644, 2019.
- [Zhou *et al.*, 2024] Muzhi Zhou, Zheng Wei, and Junxiang Liao. How can the universal disclosure of provincial-level ip geolocation change the landscape of social media analysis. *ACM SIGWEB Newsletter*, 2024(Autumn):1–9, 2024.
- [Zhu *et al.*, 2016] Yingying Zhu, Cong Yao, and Xiang Bai. Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science*, 10:19–36, 2016.