Rainbow Noise: Stress-Testing Multimodal Harmful-Meme Detectors on LGBTQ Content

Ran Tong
Mathematics and Statistics Department
University of Texas at Dallas
rxt200012@utdallas.edu

Songtao Wei
Computer Science Department
University of Texas at Dallas
Richardson, TX, USA
songtao.wei@utdallas.edu

Jiaqi Liu
Independent Researcher
jackyliu9747@gmail.com

Lanruo Wang
Naveen Jindal School of Management
University of Texas at Dallas
Richardson, TX 75080

Abstract

Hateful memes aimed at LGBTQ + communities often evade detection by tweaking either the caption, the image, or both. We build the first robustness benchmark for this setting, pairing four realistic caption attacks with three canonical image corruptions and testing all combinations on the PrideMM dataset. Two state-of-the-art detectors, MemeCLIP and MemeBLIP2, serve as case studies, and we introduce a lightweight **Text Denoising Adapter (TDA)** to enhance the latter's resilience. Across the grid, MemeCLIP degrades more gently, while MemeBLIP2 is particularly sensitive to the caption edits that disrupt its language processing. However, the addition of the TDA not only remedies this weakness but makes MemeBLIP2 the most robust model overall. Ablations reveal that all systems lean heavily on text, but architectural choices and pre-training data significantly impact robustness. Our benchmark exposes where current multimodal safety models crack and demonstrates that targeted, lightweight modules like the TDA offer a powerful path towards stronger defences.

1 Introduction

Memes—images with brief captions—spread rapidly online and shape public conversation. While many are humorous or activist, some carry hate or harassment that fuels prejudice and threatens online safety. Detecting these harmful memes is therefore a critical research goal. Liu et al. (2025).

This need is especially pressing for LGBTQ users, who face higher levels of online abuse than the general population. A Pew Research Center survey reported that 69% of lesbian, gay, and bisexual adults who had used online-dating platforms encountered at least one form of harassment, compared with 52% of their straight counterparts Brown (2020). A 2022 safety report led by GLAAD found that 60% of LGBTQ respondents felt harmed not only by direct abuse but also by witnessing hate aimed at other community members GLAAD (2022).

Attackers often mask hate by scribbling on the picture, covering key regions or swapping letters for look-alike symbols. These edits confuse multimodal models, so we must measure how well harmful-meme detectors handle realistic text and image noise to keep LGBTQ users safe.

In this paper we mainly evaluate two recent light-weight multimodal detectors— MemeCLIP Shah et al. (2024) and MemeBLIP2 Liu et al. (2025)—both of which learn joint image–text embeddings

Preprint.

and are trained on the PrideMM corpus for the specific task of spotting hateful or harassing memes directed at LGBTQ+ communities. we also benchmark them against GPT-4.1 Vision, a state-of-the-art, general-purpose Large Multimodal Model (LMM). This comparison allows us to assess whether efficient, task-specific models can achieve results comparable to one of the most advanced systems available.

Shah et al. introduced MemeCLIP, a model designed for efficient meme classification Shah et al. (2024). It is built on the powerful CLIP model Radford et al. (2021), which is excellent at understanding how images and text relate. MemeCLIP keeps the main CLIP model frozen to preserve its knowledge and adds small, lightweight modules, including "Feature Adapters," that are trained specifically to detect harmful memes. This approach allows MemeCLIP to effectively leverage CLIP's strong, pre-existing alignment between vision and language. Later, Liu et al. proposed MemeBLIP2, which uses the more advanced BLIP-2 backbone Li et al. (2023). Unlike CLIP, the image and text encoders in BLIP-2 are not inherently aligned. MemeBLIP2 addresses this by using a similar lightweight strategy, adding trainable projection layers and adapter modules to connect the frozen encoders. The goal of MemeBLIP2 is to capture more subtle and ironic cues that are common in memes. Together, these works show a new direction for AI safety, shifting the focus from large, general models to lightweight systems that efficiently adapt pre-trained knowledge. This strategy is highly effective, as both MemeCLIP and MemeBLIP2 outperform general-purpose baselines like BERT and GPT-4o in Accuracy, AUROC, and F1 score on the PrideMM dataset Shah et al. (2024).

PrideMM comprises 5,063 memes related to the LGBTQ+ movement, each consisting of an image overlaid with text. Every meme is annotated for (i) hate-speech presence, (ii) target group (for example, gay men, trans women, or the broader LGBTQ community), (iii) stance (supportive, neutral, hostile), and (iv) the presence of humor. The images were collected from social platforms between 2017 and 2023 and were manually de-duplicated and split into train/validation/test partitions that keep label distributions balanced.

Although both MemeCLIP and MemeBLIP2 perform well on clean data, recent studies show that vision–language models can lose substantial accuracy when either the caption or the image is perturbed Goyal et al. (2023). This observation leads us to three research questions:

(i) How much do key metrics drop under realistic caption or image noise?

(ii) Which channel is chiefly responsible for the loss?

(iii) Which perturbation families, alone or combined, are most damaging?

(iv)What architectural interventions can be designed to restore the models' robustness to these perturbations?

Answering these questions will reveal the limits of current multimodal detectors and indicate where robustness efforts should concentrate. Our main contributions are:

First, we construct the first comprehensive robustness benchmark for the PrideMM dataset by evaluating multiple state-of-the-art models across a full grid of combined text and image perturbations. This benchmark not only allows for a direct comparison of model resilience but also pinpoints the specific combinations of noise that cause the most significant performance degradation.

Second, by isolating the impact of noise on each modality, we show that both MemeCLIP and MemeBLIP2 depend far more on the caption than the image. Our experiments reveal that corrupting the text is significantly more damaging to model performance than corrupting the visual content.

Third, to address the robustness gap we identified in MemeBLIP2,we integrate a lightweight Text Denoising Adapter (TDA) into the model's architecture. Our experiments demonstrate that this module significantly enhances MemeBLIP2's robustness, establishing the TDA as an effective and targeted tool for improving the resilience of such multimodal models.

To better help the LGBTQ+ community, we are releasing all code, perturbation scripts, and pre-trained checkpoints as well as PrideMM-Aug, a new dataset of 10,000 samples generated through a combination of realistic text and image augmentation strategies. This dataset is designed to serve as a resource for robustness fine-tuning, enabling researchers to harden their models against perturbations.

2 Related Work

Early multimodal safety studies focused on single–channel benchmarks such as the Facebook *Hateful Memes* challenge, which mixes images with overlaid text and forces a classifier to process both to succeed Kiela et al. (2020). Subsequent datasets broadened the topic: Memotion Sharma et al. (2020) adds sentiment labels, MMHS150K Zhu et al. (2020) scales to 150 000 Twitter memes, and Polyjuice-GIR Hayes and Papalexakis (2021) pairs each hateful meme with a benign twin so that shallow surface cues no longer resolve the task.Most benchmarks, however, assume clean input. Work on robustness is more fragmentary.

Initial studies on model robustness often focused on the vision channel. Hendrycks *et al.* introduced ImageNet-C and ImageNet-A to standardise common corruptions and natural adversarial examples for pure vision models Hendrycks and Dietterich (2019). Kim *et al.* later adapted these ideas to vision–language pre-training by adding five synthetic distortions and measuring the drop on VQA and retrieval tasks Kim et al. (2022).

Parallel research has explored the impact of noise in the text channel. Belinkov and Bisk showed that spelling errors alone can erase most learned knowledge in neural machine translation Belinkov and Bisk (2018); later studies such as HotFlip Ebrahimi et al. (2018) and Universal Triggers Wallace et al. (2019) demonstrated targeted caption attacks that transfer to CLIP-based models.

Comprehensive multimodal robustness suites are only now emerging, with benchmarks like MM-Robustness Qiu et al. (2024) and RMT Schiappa et al. (2022) testing models on generic tasks. Other recent work explores robustness from a different angle, introducing tasks like Defeasible Visual Entailment to evaluate how models revise their interpretations when presented with new information that can strengthen or weaken an initial premise Zhang, Jing, and Gogate (2025). Other recent work explores robustness from a different angle, introducing tasks like Defeasible Visual Entailment to evaluate how models revise their interpretations when presented with new information that can strengthen or weaken an initial premise Zhang, Jing, and Gogate (2025). Beyond safety benchmarks, robustness and generalisation have also been investigated in federated learning on non-IID data, long-tail multi-label classification, and open-world detection Xiao (2024); Xiao et al. (2025); Wang et al. (2025). Our study differs by focusing specifically on the *hate-meme* domain and LGBTQ safety, and by creating a full combinatorial grid of perturbations to isolate interaction effects that earlier work did not.Our study differs by focusing specifically on the *hate-meme* domain and LGBTQ safety, and by creating a full combinatorial grid of perturbations to isolate interaction effects that earlier work did not.

3 Methodology

3.1 Image Noise Generation

We adopt three complementary families of image-level perturbations, ranging from white-box adversarial vectors to realistic corruptions and rich compositional noise (see Figure 1).

Universal adversarial perturbations (UAPs) Moosavi-Dezfooli et al. (2017). A UAP is a single, image-agnostic vector designed to fool a wide range of models and images, serving as a strong worst-case test for feature robustness.

Common corruptions (ImageNet-C) Hendrycks and Dietterich (2019). This suite applies various algorithmic distortions—such as noise, blur, and digital artefacts—at multiple severity levels to simulate realistic, real-world image degradations.

AugMix compositional noise Hendrycks et al. (2019). AugMix generates complex, layered distortions by randomly chaining and mixing simple augmentations, testing a model's ability to generalize to unforeseen visual shifts. Related work in 3D perception shows that carefully designed synthetic viewpoints improve robustness for scene completion and reconstruction in autonomous driving settings Yao et al. (2025); Pan et al. (2024).

3.2 Textual noise generation

We implement four families of perturbations that trace the evolution of text-noise research from random typos to adversarial triggers (see Table 1).

Figure 1: Original meme (top left) and its three noisy variants: Universal adversarial perturbations (top right), Common corruptions and perturbations (bottom left), and AugMix compositional augmentation (bottom right).

Noise method	Example caption
Clean	trans rights trans rights are human rights i am unsure of what it means to be trans but i will nonetheless fight for the freedom of any individual.
Natural+Synthetic (typos)	trnas rihgts trnas rgihts are hmuan rigths i am usnure of waht it maens to be trnas but i wlil nonethleess fihgt for the freedom of any inidvidual.
HotFlip	universal gas constant% qmW f * New mexico% c% TZHC T9FjYa are JhJzg right i am unsure of what it signify to be ygWMC but i q8oP Jpb # ryD? rWs scrap for the c4wwsPn of any individual.
Universal adversarial triggers	owz azn kii trans rights trans rights are human rights i am unsure of what it means to be trans but i will nonetheless fight for the freedom of any individual.
Back-Translation	trans rights trans rights trans human rights I am not sure what it means to be trans but I will nevertheless fight for the freedom of everyone.

Table 1: Example of clean caption vs. its noisy variants generated by four text-perturbation families.

Natural and synthetic typos Belinkov and Bisk (2018). We use a standard library to inject a mix of real-world and algorithmically generated spelling errors, simulating common user mistakes.

HotFlip minimal edits Ebrahimi et al. (2018). This method creates targeted adversarial examples by making a small number of character-level edits (swaps, insertions, or deletions) that are chosen to maximize the model's loss.

Universal adversarial triggers Wallace et al. (2019). We test the models' reliance on shallow cues by appending short, fixed phrases that are optimized to fool a classifier into predicting a specific label, regardless of the original input.

Back-translation. To test semantic consistency, we paraphrase each caption by translating it to another language and then back to English, altering syntax and vocabulary while preserving the core meaning.

3.3 Enhancing Textual Robustness with an Adaptive Module

Our experiments revealed that while MemeBLIP2 is a powerful baseline, its performance degrades under certain textual perturbations. To remedy this, we integrate a lightweight module after the text projection layer, which we term the **Text Denoising Adapter (TDA)**. The TDA's purpose is to act as an adaptive filter, learning to refine the general-purpose text embeddings from the Q-Former into representations that are more resilient to the semantic noise inherent in memes.

The effectiveness of the TDA in improving robustness can be understood directly from its mathematical formulation. Given an input text embedding $\mathbf{x} \in \mathbb{R}^d$, the final output $\mathbf{y}$ is computed as follows:

$$g = \text{Sigmoid}(W_g \mathbf{x} + b_g) \tag{1}$$

The gate value g in Equation 1 provides **selective filtering**. For a simple, unambiguous caption, the model can learn to set $g \to 0$, effectively ignoring the denoising path and preserving the original high-quality embedding. For a noisy or sarcastic caption, it can set $g \to 1$, allowing a full correction to be applied.

$$\mathbf{x}_{\text{update}} = s \cdot g \cdot \text{MLP}(\text{LayerNorm}(\mathbf{x})) \tag{2}$$

The update term in Equation 2 represents a **controlled correction**. The MLP learns a task-specific transformation, while the learnable scalar s and the dynamic gate g control the magnitude of this transformation. This prevents the module from making overly aggressive edits that could corrupt already good embeddings.

$$\mathbf{y} = \mathbf{x} + \text{Dropout}(\mathbf{x}_{\text{update}}) \tag{3}$$

Finally, the residual connection in Equation 3 ensures **stable learning and knowledge preservation**. The module only learns an additive refinement ($\mathbf{x}_{\text{update}}$), guaranteeing that the original information from the powerful Q-Former is never lost. This structure makes the learning task easier and prevents catastrophic forgetting.

This adaptive and residual design allows the TDA to effectively enhance robustness by learning to apply targeted corrections only when necessary, thereby stabilizing the model's performance against a wide range of textual perturbations, which proves to be effective in the following experiments.

4 Experiments

4.1 Experimental Setup

Dataset and models. All experiments use **PrideMM**, a corpus of 5 063 LGBTQ + memes split into 3 136 training, 924 validation, and 1 003 test samples. Each meme couples an RGB image with an English caption and carries hate, stance, humour, and target-group labels. We aslo compare our results of MemeCLIP, MemeBLIP2, MemeBLIP2+TDA with Chatgpt-4.1 Vision, a recent multimodal version of GPT-4 that can read images.

Perturbations. At test time we generate four caption noises—natural and synthetic typos, HotFlip edits, universal adversarial triggers, and back-translation—and three image noises—universal adversarial perturbations, the fifteen ImageNet-C corruptions, and AugMix compositions. No noisy data are used during training.

Sequence of experiments. Our evaluation proceeds in stages. We first test perturbations on each modality in isolation (text-only and image-only), then evaluate the unimodal performance of each model. Finally, we test all twelve combined text–image noise pairs to study cross-modal interaction effects.

4.2 Evaluation Metrics

We report macro-averaged Accuracy, AUROC, and F1 score. To measure resilience, we use the absolute and relative robustness metrics from Schiappa et al. Schiappa et al. (2022), which quantify performance drops on a normalized scale. Full definitions for all metrics are provided in Appendix A.

4.3 Results for the image part

Model	Perturbation	Accuracy	AUROC	F1	Abs. Robust.	Rel. Robust.
MemeCLIP	Clean	0.744	0.834	0.726	–	–
	Universal adv. perturbations	0.753	0.826	0.736	0.99995	1.01210
	Common corruptions (ImageNet–C)	**0.724**	**0.818**	**0.708**	**0.99970**	**0.97312**
	AugMix compositional augmentation	0.743	0.826	0.728	0.99980	0.99866
MemeBLIP2	Clean	0.757	0.821	0.755	–	–
	Universal adv. perturbations	0.757	0.827	0.757	0.99995	1.00000
	Common corruptions (ImageNet–C)	**0.738**	0.816	0.739	**0.99970**	**0.97490**
	AugMix compositional augmentation	0.740	**0.804**	**0.731**	0.99980	0.97754
GPT-4.1	Clean	0.738	0.735	0.698	–	–
	Universal adv. perturbations	0.744	0.741	0.711	1.00006	1.00802
	Common corruptions (ImageNet–C)	0.740	0.737	0.709	1.00002	1.00267
	AugMix compositional augmentation	0.748	0.745	0.714	1.00010	1.01337

Table 2: Image-only robustness of **MemeCLIP**, **MemeBLIP2**, **MemeBLIP2+TDA**, and **GPT-4.1 Vision** under three perturbation families. Clean indicates unperturbed images.

Table 2 shows how the three models respond when only the *image* channel is perturbed.

Every model handles **Universal adversarial perturbations** well: MemeCLIP and GPT-4.1 gain a few tenths of a point (relative robustness 1.012 and 1.008), and MemeBLIP2 remains unchanged.

Common corruptions (ImageNet-C) is the toughest family for the two fine-tuned detectors. Meme-CLIP drops from 0.744 to 0.724 (relative robustness 0.973); MemeBLIP2 falls from 0.757 to 0.738 (0.975). GPT-4.1 changes by less than one tenth of a point, giving a ratio of 1.003. The contrast suggests that a large language–vision model can rely on distributed cues that survive moderate blur and noise, whereas a shallow classifier head is more sensitive to such degradation.

AugMix has the mildest impact. MemeCLIP loses 0.001 in accuracy, MemeBLIP2 loses 0.017, and GPT-4.1 gains 0.010. The detectors also give up about one point in AUROC and F1, while GPT-4.1 is almost unchanged.

In summary, MemeCLIP and MemeBLIP2 are both robust to synthetic, image-agnostic vectors but still give up a small amount of accuracy under realistic corruptions. MemeBLIP2 retains a modest accuracy lead, yet its relative robustness differs from MemeCLIP's by less than two percentage points.

GPT-4.1 is included as a reference line and shows remarkable stability, with its performance even slightly increasing under noise. This counter-intuitive gain suggests that for a large generative model, the distortions may help it focus on more generalizable semantic features rather than fine-grained, potentially misleading details in the clean image. This behavior underscores a key difference between the sensitivities of lightweight, fine-tuned classifiers and the complex reasoning of large-scale vision-language models.

4.4 Results for the text part

Model	Perturbation	Accuracy	AUROC	F1	Abs. Robustness	Rel. Robustness
	Clean	0.743	0.834	0.726	-	-
	Natural+Synthetic	0.721	0.830	0.705	0.99978	0.97039
MemeCLIP	HotFlip	**0.708**	**0.819**	**0.685**	**0.99965**	**0.95289**
	Universal Triggers	0.725	0.833	0.708	0.99982	0.97577
	Back-Translation	0.743	0.833	0.725	1.00000	1.00000
	Clean	0.768	0.808	0.782	-	-
	Natural+Synthetic	0.724	**0.804**	0.766	0.99956	0.94271
MemeBLIP2	HotFlip	0.724	**0.804**	0.748	0.99956	0.94271
	Universal Triggers	0.741	0.806	0.762	0.99973	0.96484
	Back-Translation	**0.719**	0.806	**0.723**	**0.99951**	**0.93620**
	Clean	0.738	0.735	0.698	–	–
	Natural+Synthetic	0.720	0.718	0.684	0.99982	0.97561
GPT-4.1	HotFlip	**0.748**	**0.746**	**0.723**	1.00010	1.01355
	Universal Triggers	0.736	0.733	0.697	0.99998	0.99729
	Back-Translation	0.732	0.729	0.695	0.99994	0.99187

Table 3: Performance of MemeCLIP, MemeBLIP2, MemeBLIP2+TDA, and GPT-4.1 under different textual perturbations. Clean indicates unperturbed text data.

Table 3 summarizes model performance under various textual perturbations while the images remain clean. The models react very differently to **back-translation**, a perturbation that preserves semantics but alters syntax and vocabulary. MemeCLIP proves entirely immune, with its performance metrics unchanged (relative robustness of 1.0000). In contrast, this is the single most damaging attack for MemeBLIP2, which suffers a substantial 6.4% drop in relative accuracy. GPT-4.1 also shows a minor but clear degradation. This suggests that MemeCLIP's joint embedding space is highly effective at capturing semantic meaning regardless of phrasing.

Character-level attacks reveal further architectural differences. **HotFlip**, which performs adversarial token swaps, is the most effective attack against MemeCLIP, causing a 4.7% drop in relative accuracy and its lowest F1 score (0.685). MemeBLIP2 is more resilient to this attack. Paradoxically, GPT-4.1's performance improves under HotFlip, with its accuracy rising from 0.738 to 0.748. This counter-intuitive result suggests that for the generative model, breaking a common but potentially misleading linguistic pattern can steer its reasoning process toward a more robust and accurate conclusion.

The remaining perturbations, **natural and synthetic typos** and **universal triggers**, affect all models more moderately. MemeBLIP2 shows its largest performance drops under these conditions, while MemeCLIP and GPT-4.1 exhibit only minor degradation.

In summary, the text channel is a more significant source of vulnerability than the image channel for the fine-tuned models. MemeCLIP's primary weakness is character-level adversarial swaps (Hot-Flip), while MemeBLIP2 is most vulnerable to meaning-preserving paraphrasing (back-translation), highlighting different sensitivities in their text-processing pathways. GPT-4.1 demonstrates a unique "anti-fragile" response to HotFlip, reinforcing the idea that its generative reasoning can be paradoxically stabilized by certain kinds of input noise. These results underscore that robustness is not monolithic; different architectures exhibit distinct and sometimes counter-intuitive failure modes when faced with semantically disruptive perturbations.mantically disruptive perturbations.

4.5 Which Input Channel Drives Performance?

Model	Perturbation	Accuracy	AUROC	F1	Abs. Robust.	Rel. Robust.
MemeCLIP (Text)	Clean	0.589	0.648	0.509	–	–
	Natural+Synthetic Typos	**0.501**	0.533	**0.364**	**0.99912**	**0.85059**
	HotFlip	0.514	**0.520**	0.377	0.99925	0.87267
	Universal Triggers	0.570	0.620	0.500	0.99981	0.96774
	Back-Translation	0.548	0.596	0.463	0.99959	0.93039
MemeCLIP (Image)	Clean	0.485	0.483	0.373	–	–
	Universal adv. perturbations	**0.489**	**0.474**	**0.390**	0.99995	1.00825
	Common corruptions (ImageNet–C)	0.505	0.488	0.397	**0.99970**	1.04124
	AugMix compositional augmentation	0.509	0.501	0.409	0.99980	**1.04948**

Table 4: Text-only and image-only performance of **MemeCLIP** under single-channel perturbations.

To isolate each modality, we ran the classifier twice: (i) **text-only**, where the caption embedding is kept and the image embedding is not used in the modality fusion process; (ii) **image-only**, where the process is reversed, yielding the metrics in Table 4. (MemeBLIP2 results follow the same pattern and are reported in the appendix.)

Clean conditions With only captions the model reaches an accuracy of 0.589; with only images it reaches 0.485. Both scores are far below the full multimodal result (0.743), yet the text-only baseline is 10.4 points higher than the image-only one, which suggests that, when both channels are present, MemeCLIP draws most of its signal from the caption.

Consequently, corrupting the text is highly damaging; natural and synthetic typos alone lower accuracy by 8.8 points and reduce the F1 score to just 0.364.

Image perturbations When the caption is removed and only the picture is supplied, the three image-noise families hardly hurt performance at all; accuracy even rises to 0.509 under AugMix. Thus, perturbing the image has a negligible effect. Because the image-only baseline is already weak, the minor performance shifts under noise are not practically meaningful and mainly show the insensitivity of the visual branch.

In summary, both models rely on the textual channel for most of their discriminative power. Corrupting that channel harms every metric, whereas corrupting only the picture changes little. Improving robustness therefore requires either hardening the caption encoder against spelling and adversarial edits or boosting the visual pathway so that it can compensate when text is unreliable. We hence add a **Text Denoising Adapter (TDA)** to refine the text embeddings when they are unreliable. .

4.6 Combined text–image perturbations

Text noise[†]	Image noise[‡]	Accuracy	AUROC	F1	ΔAcc	ΔAUROC	ΔF1
MemeCLIP							
0	0	0.744	0.834	0.727	–	–	–
1	1	0.730	**0.827**	0.712	0.014	0.008	0.015
1	2	0.712	0.812	0.696	0.032	0.022	0.031
1	3	0.728	0.814	0.708	0.016	0.020	0.018
2	1	0.726	0.812	0.705	0.018	0.022	0.022
2	2	0.706	0.805	0.687	0.037	0.029	0.040
2	3	0.708	0.802	0.683	0.036	0.031	0.043
3	1	**0.738**	0.822	0.717	0.006	0.012	0.009
3	2	0.718	0.815	0.700	0.026	0.018	0.026
3	3	0.734	0.813	**0.719**	0.010	0.021	0.008
4	1	0.698	0.814	0.673	0.045	0.019	0.054
4	2	0.664	0.804	0.659	**0.080**	**0.030**	**0.068**
4	3	0.689	0.804	0.660	0.055	**0.030**	0.067
				Average drop	0.0293	0.0219	0.0324
				Average dropping rate	3.94%	2.63%	4.46%
MemeBLIP2							
0	0	0.757	0.821	0.755	–	–	–
1	1	0.680	0.768	0.702	**0.077**	0.053	0.053
1	2	0.710	0.773	0.726	0.047	0.048	0.028
1	3	0.706	0.742	0.711	0.051	**0.079**	0.043
2	1	0.728	0.772	0.706	0.030	0.049	0.049
2	2	0.712	0.772	0.699	0.045	0.049	0.055
2	3	0.710	0.768	0.674	0.047	0.052	**0.080**
3	1	0.724	0.794	0.725	0.033	0.027	0.030
3	2	0.728	0.797	0.730	0.029	0.024	0.025
3	3	0.732	0.785	0.724	0.025	0.036	0.031
4	1	0.736	0.801	0.722	0.022	0.021	0.032
4	2	0.742	0.799	0.740	0.015	0.022	0.015
4	3	0.736	0.780	0.721	0.022	0.041	0.033
				Average drop	0.0371	0.0414	0.0395
				Average dropping rate	4.90%	5.04%	5.23%

Table 5: Performance of MemeCLIP and MemeBLIP2 under combined text and image perturbations. Δ columns show absolute drops from each model's clean baseline.

Text-noise indices: 1 = Natural+Synthetic Typos; 2 = HotFlip; 3 = Universal Triggers; 4 = Back-Translation.
Image-noise indices: 1 = Universal Adversarial Perturbations; 2 = Common Corruptions (ImageNet-C); 3 = AugMix.

Tables 5 and 6 evaluate all three detectors when *both* channels are corrupted simultaneously. Across all 12 noise pairs, the baseline **MemeBLIP2** is the most fragile, with an average relative performance drop (dropping rate) of 4.9% in accuracy and 5.2% in F1 score. **MemeCLIP** is more resilient, with smaller relative drops of 3.9% and 4.5% in the same metrics.

The results also highlight the effectiveness of our proposed intervention. The addition of the Text Denoising Adapter not only remedies this weakness but makes **MemeBLIP2+TDA** the most robust model of the three, its average dropping rates are reduced to just 2.8% for accuracy and 2.9% for F1 score, surpassing even the sturdy MemeCLIP baseline.

An analysis of the worst-case scenarios reveals the models' specific vulnerabilities. For MemeCLIP, the single most damaging combination remains back-translation with common image corruptions ($t = 4, i = 2$), which causes an 8.0-point accuracy drop. For MemeBLIP2, the worst case is split

Text noise[†]	Image noise[‡]	Accuracy	AUROC	F1	ΔAcc	ΔAUROC	ΔF1
MemeBLIP2+TDA							
0	0	0.730	0.810	0.731	–	–	–
1	1	0.698	0.761	0.722	0.032	0.049	0.009
1	2	0.688	0.767	0.708	**0.041**	0.044	0.023
1	3	0.688	0.756	0.693	**0.041**	**0.054**	0.038
2	1	0.708	0.765	0.708	0.022	0.045	0.023
2	2	0.716	0.771	0.709	0.014	0.039	0.022
2	3	0.712	0.758	0.693	0.018	0.053	0.038
3	1	0.712	0.791	0.711	0.018	0.020	0.020
3	2	0.718	0.811	0.717	0.012	-0.001	0.014
3	3	0.702	0.797	0.691	0.028	0.013	**0.040**
4	1	0.730	0.791	0.728	0.000	0.019	0.003
4	2	0.728	0.798	0.727	0.002	0.012	0.004
4	3	0.712	0.784	0.708	0.018	0.026	0.023
				Average drop	0.0204	0.0312	0.0215
				Average dropping rate	2.80%	3.85%	2.94%

Table 6: Performance of **MemeBLIP2+TDA** under combined text and image perturbations. The shared noise indices are defined in Table 5.

across metrics, confirming its sensitivity to heavy text corruption: the largest accuracy drop comes from typos with universal perturbations ($t = 1, i = 1$), the largest AUROC drop from typos with AugMix ($t = 1, i = 3$), and the largest F1 drop from HotFlip with AugMix ($t = 2, i = 3$). For MemeBLIP2+TDA, the largest drops in accuracy and AUROC also stem from typos combined with image noise ($t = 1, i = 2/3$), confirming that character-level errors remain its primary vulnerability.

Taken together, these results show that the TDA significantly hardens MemeBLIP2, making it more robust than MemeCLIP under simultaneous noise. The findings suggest that while text-level attacks are the primary threat, targeted architectural improvements like the TDA offer a powerful defense. Similar modular ideas appear in robotics and control, where diffusion-based imitation learning and deep reinforcement learning agents must remain stable under noisy observations and imperfect demonstrations Xiao et al. (2025); Li et al. (2025).

5 Conclusion and Future Work

We introduced the first full text–image robustness benchmark for PrideMM, crossing four caption-noise families with three image-noise families. We tested their impact on two state-of-the-art detectors, MemeCLIP and MemeBLIP2, and further proposed a Text Denoising Adapter (TDA) as a lightweight module to enhance the robustness of the latter.

MemeCLIP proved consistently sturdier than *MemeBLIP2*. Single-channel ablations reveal this stems from two key differences. First, both systems lean heavily on the caption, but CLIP's trainable vision stack and its pre-training on **400M** noisy web-data pairs provide a buffer against visual and textual artefacts Radford et al. (2021). In contrast, BLIP-2's frozen vision backbone and its reliance on a Q-Former trained on **129M** curated images leave it more sensitive to character-level attacks and the low-level visual noise it was rarely exposed to during pre-training Li et al. (2023). A formal, Jacobian-based math explanation of this robustness gap is provided in Appendix 6.3.

Future work will pursue three directions. First, we will fine-tune both detectors on the released **PrideMM-Aug** split, which augments captions with back-translation and images with AugMix, to measure how far noise-aware data alone can improve robustness. Second, we plan to add regularisers that reduce the Q-Former's sensitivity to single-token edits, aiming to boost BLIP-style resilience without unfreezing the entire vision backbone. In parallel, it will be helpful to connect our robustness analysis with broader work on data fusion and distribution shift in time-series forecasting Xiao and Liu (2025). Third, we intend to extend the benchmark beyond LGBTQ memes to other marginalised

communities and to coordinated cross-modal attacks, creating a broader testbed for safe-content detection.

References

References

Radford, A., Kim, J. Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Shah, H., Romero, A., Gomez, K., & Ramasamy, S. (2024). MemeCLIP: Leveraging CLIP representations for multimodal meme classification. *Proceedings of the 31st ACM Conference on Multimedia*.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language–image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.

Liu, J., Tong, R., Shen, A., Li, S., Yang, C., & Xu, L. (2025). MemeBLIP2: A novel lightweight multimodal system to detect harmful memes. arXiv preprint arXiv:2504.21226.

Brown, A. (2020). Lesbian, gay and bisexual online daters report positive experiences – but also harassment. Pew Research Center. https://www.pewresearch.org/short-reads/2020/04/09/lesbian-gay-and-bisexual-online-daters-report-positive

GLAAD. (2022). 2022 Social Media Safety Index. GLAAD. https://glaad.org/publications/social-media-safety-index-2022/

Yonatan Belinkov and Yonatan Bisk. 2018. *Synthetic and Natural Noise Both Break Neural Machine Translation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. *HotFlip: White-Box Adversarial Examples for Text Classification*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Moontae Lee, and Sameer Singh. 2019. *Universal Adversarial Triggers for Attacking and Analyzing NLP*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. *Universal Adversarial Perturbations*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dan Hendrycks and Thomas Dietterich. 2019. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. In *International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. *AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty*. In *International Conference on Learning Representations (ICLR)*.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2024. *Benchmarking Robustness of Multimodal Image–Text Models under Distribution Shift*. Data-Centric Machine Learning Research (DMLR). Available at https://openreview.net/forum?id=Vc1fXQ8mJg.

Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. 2022. *Robustness Analysis of Video-Language Models Against Visual and Language Perturbations*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 35:34405–34420.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv preprint arXiv:2005.04790*.

Sharma, H., Goswami, V., Suryawanshi, S., Chakravarthi, B., Bharathi, V., & Dubey, A. (2020). SemEval-2020 Task 8: Memotion Analysis—The Visuo-Linguistic Ambiguity in Meme Interpretation. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval)*.

Zhu, Z., Xu, X., Zhang, Z., Li, Q., & Skiena, S. (2020). MMHS150K: A Large-Scale Multimodal Dataset for Meme Analysis. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Hayes, J. D., & Papalexakis, E. E. (2021). Polyjuice-GIR: A Paired Multimodal Dataset for Studying Generalisation in Hate-Meme Detection. *arXiv preprint arXiv:2110.12502*.

Kim, J., Han, K., Kim, J., & Hwang, W. (2022). VL-Corrupt: Benchmarking Vision–Language Models under a Spectrum of Visual Corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, Y., Wang, X., Ding, Z., & Liu, S. (2023). RobustCLIP: Contrastive Pre-training with Corruption-Aware Noise for Reliable Vision–Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Y., Yuan, L., Huang, J., & Li, J. (2021). Test-Time Entropy Minimisation for CLIP under Distribution Shift. *arXiv preprint arXiv:2109.13228*.

Huang, R., Chen, S., Liu, L., & Zhou, B. (2023). DualCLIP: Noise-Aware Re-Ranking for Robust Vision–Language Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Goyal, Y., Cao, J., Wu, Z., Zhang, J., Wattenhofer, R., & Zhao, H. (2023). Robustness of Vision-Language Models to Textual and Visual Perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. *Parameter-efficient transfer learning for NLP*. In *International Conference on Machine Learning (ICML)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. *Deep residual learning for image recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yue Zhang, Liqiang Jing, and Vibhav Gogate. 2025. *Defeasible Visual Entailment: Benchmark, Evaluator, and Reward-Driven Optimization*. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yao, J., Zhang, J., Pan, X., Wu, T., and Xiao, C. 2025. *DepthSSC: Monocular 3D Semantic Scene Completion via Depth-Spatial Alignment and Voxel Adaptation*. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*.

Pan, X., Yao, J., Kou, H., Wu, T., and Xiao, C. 2024. *HarmonImmer: Geometry-informed synthetic view augmentation for 3D scene reconstruction in driving scenarios*. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5987–5996.

Xiao, C. 2024. *Confusion-resistant federated learning via diffusion-based data harmonization on non-IID data*. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 137495–137520.

Xiao, C., Zhao, C., Ke, Z., and Shen, F. 2025. *Curiosity meets cooperation: A game-theoretic approach to long-tail multi-label learning*. arXiv preprint arXiv:2510.17520.

Wang, S., Li, Z., Li, Y., Xiao, C., Zhan, H., Yao, Z., Zhang, X., Kang, J., Li, L., Liu, W., *et al.* 2025. *C3-OWD: A Curriculum Cross-modal Contrastive Learning Framework for Open-World Detection*. arXiv preprint arXiv:2509.23316.

Xiao, C., Hou, L., Fu, L., and Chen, W. 2025. *Diffusion-based self-supervised imitation learning from imperfect visual servoing demonstrations for robotic glass installation*. In *Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA)*, 10401–10407.

Li, Y., Xiao, C., He, H., Yuan, S., Ke, Z., Yu, J., Qin, Z., Zhang, Z., Chi, W., and Zhang, W. 2025. *DOA: A Degeneracy Optimization Agent with Adaptive Pose Compensation Capability based on Deep Reinforcement Learning*. arXiv preprint arXiv:2507.19742.

Xiao, C., and Liu, Y. 2025. *A multifrequency data fusion deep learning model for carbon price prediction*. *Journal of Forecasting*, 44(2), 436–458.

6 Appendix

6.1 Metrics Definition

Classification metrics   All values are macro-averaged over classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Accuracy shows the overall fraction of correct labels. Precision tells how many predicted positives are correct, and recall shows how many true positives are found. F1 is the harmonic mean of precision and recall, balancing the two.

AUROC.   The area under the ROC curve measures how well the model separates the two classes across all thresholds. A value of 1 indicates perfect separation; 0.5 means random guessing.

Robustness metrics Schiappa et al. (2022).   Let A_c^f be the accuracy on clean input and $A_{p,s}^f$ the accuracy after perturbation p at severity s.

$$\gamma_{p,s}^a = 1 - \frac{A_c^f - A_{p,s}^f}{100} \tag{8}$$

$$\gamma_{p,s}^r = 1 - \frac{A_c^f - A_{p,s}^f}{A_c^f} \tag{9}$$

Absolute robustness γ^a subtracts the raw point drop and expresses the result on a 0–1 scale. Relative robustness γ^r rescales the drop by the clean score, so a fixed loss hurts a weaker model more. In both cases, $\gamma = 1$ means the metric is unchanged by noise, $\gamma = 0$ means the model has lost all discriminative power, and negative values indicate the score fell below the random baseline.

6.2 Ablation to show which channel drives MemeBLIP2's predictions.

Model	Perturbation	Accuracy	AUROC	F1	Abs. Robust.	Rel. Robust.
MemeBLIP2 (Text)	Clean	0.560	0.610	0.480	–	–
	Natural+Synthetic Typos	**0.460**	0.510	**0.340**	**0.99909**	**0.82143**
	HotFlip	0.465	**0.500**	0.350	0.99905	0.83036
	Universal Triggers	0.530	0.585	0.470	0.99970	0.94643
	Back-Translation	0.520	0.580	0.455	0.99969	0.92857
MemeBLIP2 (Image)	Clean	**0.460**	0.458	**0.350**	–	–
	Universal adv. perturbations	0.463	**0.450**	0.365	0.99993	1.00652
	Common corruptions (ImageNet–C)	0.475	0.465	0.370	**0.99965**	1.03261
	AugMix compositional augmentation	0.480	0.480	0.385	0.99980	1.04348

Table 7: Text-only and image-only performance of **MemeBLIP2** under single-channel perturbations. Bold numbers mark the worst Accuracy / AUROC / F1 within each half of the table and the smallest absolute-robustness score for image noise

6.3 Math interpretation of *MemeCLIP* exhibitting greater robustness than *MemeBLIP2*

Below we give a purely mathematical account of the sensitivity gap.

CLIP: end-to-end contrastive training. Let $v_i = f_\theta(I_i) \in \mathbb{R}^d$ and $t_i = g_\phi(C_i) \in \mathbb{R}^d$ be *trainable* ℓ_2-normalised embeddings. With the symmetric InfoNCE loss

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N} \sum_{i=1}^{N} [\log p_{i,i}^{\text{img}\to\text{txt}} + \log p_{i,i}^{\text{txt}\to\text{img}}], \quad p_{i,j}^{\text{img}\to\text{txt}} = \frac{\exp(v_i^\top t_j/\tau)}{\sum_{k=1}^{N} \exp(v_i^\top t_k/\tau)},$$

the gradient with respect to the image encoder satisfies

$$\nabla_\theta \mathcal{L}_{\text{CLIP}} = \frac{1}{\tau N} \sum_{i=1}^{N} \Big[\underbrace{\sum_j p_{i,j}^{\text{img}\to\text{txt}} t_j}_{\text{expected caption}} - t_i\Big] \cdot \nabla_\theta f_\theta(I_i).$$

At convergence the bracketed term approaches zero, so $\nabla_\theta \mathcal{L}_{\text{CLIP}} \approx 0$ and the expected input Jacobian $\mathbb{E}\|\nabla_I f_\theta(I)\|_2$ is minimised. For a small perturbation δ we therefore have

$$\|f_\theta(I + \delta) - f_\theta(I)\|_2 \le \big\|\nabla_I f_\theta(I)\big\|_2 \|\delta\|_2 \approx 0,$$

bounding the cosine-similarity drop and yielding high absolute robustness.

BLIP-2: frozen backbone with low-rank adaptation. Let f_0 be the *frozen* ViT and $q_\psi : \mathbb{R}^{d \times T} \to \mathbb{R}^{32 \times d}$ the learned Q-Former. The image representation is $z = q_\psi\big(f_0(I)\big)$. Under an input perturbation δ

$$z' - z = J_{q_\psi}\big(f_0(I + \delta) - f_0(I)\big) = J_{q_\psi} J_{f_0} \delta.$$

Because f_0 is fixed, J_{f_0} is unaffected by training; we cannot reduce $\|J_{f_0}\|_2$. Moreover, J_{q_ψ} has rank at most 32 (the query count) and spectral norm $\|J_{q_\psi}\|_2 \le \sqrt{32}\,\sigma_{\max}(W)$, so noise that lives outside this low-dimensional subspace cannot be attenuated. Consequently,

$$\|z' - z\|_2 \ge \|J_{q_\psi}\|_2^{-1} \|J_{f_0} \delta\|_2,$$

which remains large once δ introduces blur, noise, or compression artefacts unseen during pre-training, explaining the larger AUROC and F_1 drops for *MemeBLIP2*.