UniSafe: Modality-Agnostic Hateful Content Detection via Shared-Space Projection

Siam Shibly Antar
siam.antar@mcgill.ca
School of Computer Science, McGill University
Montreal, Canada

Syem Shibly Ador
syemshibly.ador@students.mq.edu.au
School of Computer Science, Macquarie University
Sydney, Australia

Steven H. H. Ding
steven.h.ding@mcgill.ca
School of Information Studies, McGill University
Montreal, Canada

Benjamin C. M. Fung
ben.fung@mcgill.ca
School of Information Studies, McGill University
Montreal, Canada

Abstract

Hateful content on social platforms is often conveyed through *multimodal memes*, where meaning emerges from image–text composition. While multimodal fusion models can be accurate, many are brittle when a modality is unavailable at inference (e.g., missing image due to broken media links, or missing text due to upstream extraction or logging failures), requiring separate fallback models or routing logic. We present **the UniSafe framework**, a *modality-agnostic* approach that projects image and text evidence into a shared *safety embedding space* using frozen foundation encoders and lightweight trainable projectors. The UniSafe framework applies a single classifier to an order-invariant aggregation of the available modalities, enabling inference on *any non-empty subset* without retraining. On the Hateful Memes benchmark [5] (dev split), UniSafe matches late fusion (AUROC 0.6865 ± 0.0026 vs. 0.6853 ± 0.0016) while supporting missing-modality inference with the *same checkpoint*. Under modality failures, UniSafe is competitive with a fusion+fallback deployment (AUROC 0.6509 on image-only and 0.6489 on text-only). Ablations indicate that *modality dropout* is the primary driver of robustness, while the alignment regularizer is optional at this scale.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Natural language processing**; *Computer vision*; • **Information systems** → *Social networking sites*; • **Security and privacy** → *Human and societal aspects of security and privacy*.

Keywords

multimodal learning; hateful content detection; memes; missing-modality robustness; foundation models; any-subset inference

ACM Reference Format:
Siam Shibly Antar, Syem Shibly Ador, Steven H. H. Ding, and Benjamin C. M. Fung. 2026. UniSafe: Modality-Agnostic Hateful Content Detection via Shared-Space Projection. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3774905.3795455

WWW Companion '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2308-7/2026/04
https://doi.org/10.1145/3774905.3795455

1 Introduction

Hateful content on social media has evolved from simple text slurs to complex multimodal compositions, particularly *memes* [2, 5, 8, 10], where toxic meaning often arises from the interaction between image and text. In the Hateful Memes benchmark [5], many samples are *unimodal-hard*: the text and image may appear benign in isolation, but their combination constitutes hate. While recent vision-language models achieve high accuracy by modeling dense cross-modal interactions, they typically operate under a strict assumption: *complete input availability*.

Problem: fragility of fusion. In production systems, inputs are rarely perfect. Images may fail to load (broken links, storage outages), and text extraction or logging can fail. Standard fusion architectures (e.g., VisualBERT [6]) treat the input as a rigid pair (I, T) and degrade when a modality is missing, often requiring separate unimodal checkpoints plus routing logic.

Scope. We focus on **static image–text meme posts**. We use the Hateful Memes benchmark [5] and do not introduce a separate OCR pipeline, utilizing the provided text modality directly.

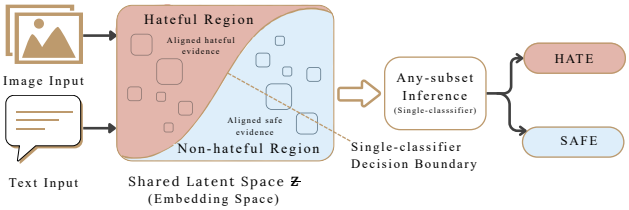

Figure 1: Safety embedding intuition. The UniSafe framework projects image and text evidence into a shared latent space $\mathcal{Z}$. A single classifier separates hateful vs. safe regions and supports *any-subset* inference.

Our approach: shared-space projection for any-subset inference. We treat multimodal evidence as a variable-sized *set* rather than a fixed paired input (Figure 1). The UniSafe framework projects each available modality into a shared *safety embedding space* and then applies a permutation-invariant aggregation before classification. This yields a simple inference contract:

Input: $\mathcal{M}_x \subseteq \{\texttt{img}, \texttt{text}\}$, $\mathcal{M}_x \neq \emptyset$ $\Rightarrow$

Output: $p(y{=}1 \mid \mathcal{M}_x)$ from one checkpoint.

Contributions.

- **Modality-agnostic shared-space projection:** lightweight projectors map frozen foundation embeddings into a common safety space optimized for hate prediction.
- **Any-subset inference with one checkpoint:** permutation-invariant aggregation (centroid pooling) supports {img}, {text}, or {img,text} without retraining.
- **Robustness via Modality Dropout:** We demonstrate that training with stochastic modality dropout enables competitive performance on incomplete inputs without requiring complex fallback routing. Our ablations further reveal that this dropout strategy is the primary driver of robustness, rendering alignment regularization a supplementary rather than necessary component.

2 Related Work and Positioning

Multi-modal hate detection. The Hateful Memes dataset [5] shifted attention from unimodal NLP to vision-language reasoning. State-of-the-art approaches often use large cross-modal transformers such as VisualBERT [6], UNITER [3], LXMERT [12], or ViLBERT [7]. While accurate, these models are computationally intensive and often require complete inputs at inference.

Missing modality robustness. Handling missing data has a long history, including zero-imputation or generative imputation, both of which can introduce noise or latency [1]. Modality dropout [9] trains models to tolerate absent signals.

Our approach: embedding-first + set aggregation. The UniSafe framework adopts an embedding-first pipeline with frozen encoders (OpenCLIP [4, 11]) and a set-function classifier. Centroid pooling is permutation-invariant and fits the standard Deep Sets formulation for variable-sized evidence [13]. The key design point is that shared-space projection (and optional regularization) yields a single decision head that can operate on any non-empty subset of modalities.

3 UniSafe Framework: Shared-Space Projection for Any-Subset Inference

We treat multimodal evidence as a variable-sized set rather than a fixed paired input (Figure 1).

Modalities as a set. Each sample contains a set of observed modalities:

$$x = \{m\}_{m \in \mathcal{M}_x}, \quad y \in \{0,1\}, \qquad \mathcal{M}_x \subseteq \{\texttt{img}, \texttt{text}\}.$$

We require $\mathcal{M}_x$ to be non-empty at inference.

Frozen encoders and trainable projectors. We use frozen foundation encoders to obtain unimodal embeddings:

- **Vision encoder** $E_{\texttt{img}}$: CLIP-style ViT-B/16 producing $v \in \mathbb{R}^{512}$.
- **Text encoder** $E_{\texttt{text}}$: paired CLIP-style text tower producing $t \in \mathbb{R}^{512}$.

We L2-normalize both embeddings before projection. Each modality is mapped into a shared space $\mathcal{Z}$:

$$z_{\texttt{img}} = P_{\texttt{img}}(v) \in \mathbb{R}^D, \qquad z_{\texttt{text}} = P_{\texttt{text}}(t) \in \mathbb{R}^D.$$

Projector heads. $P_{\texttt{img}}$ and $P_{\texttt{text}}$ are two independent **2-layer** MLPs of the form *Linear–GELU–Dropout–Linear*, with hidden width 512 and output dimension D=256.

Order-invariant aggregation and classification. UniSafe aggregates available projections using a permutation invariant centroid [13]:

$$z_{\text{final}} = \frac{1}{|\mathcal{M}_x|} \sum_{m \in \mathcal{M}_x} z_m, \qquad \hat{y} = \sigma(W z_{\text{final}} + b).$$

This yields a fixed-size representation regardless of whether one or both modalities are present.

Why centroid pooling? We use centroid pooling as a minimal Deep Sets instantiation: it is permutation-invariant, parameter-free, and yields a fixed-size representation regardless of whether one or both modalities are present. This choice avoids introducing additional fusion parameters that may implicitly assume complete inputs, and keeps the framework's any-subset inference contract explicit. While richer set functions (e.g., attention-based pooling) are possible, centroid aggregation provides a strong baseline that isolates the effect of shared-space projection and modality dropout on missing-modality robustness.

Training objective: modality dropout + optional alignment.

- ***Modality dropout.*** During training we randomly drop modalities with probability p while ensuring at least one remains, explicitly training the framework to operate under partial evidence [9].
- ***Alignment regularization (optional).*** When both modalities are present (not dropped), we can encourage agreement in $\mathcal{Z}$:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{align}}, \qquad \mathcal{L}_{\text{align}} = \|z_{\texttt{img}} - z_{\texttt{text}}\|_2^2.$$

We prioritize modality dropout as the critical factor for robustness (see Section 5), whereas alignment plays a secondary role.

4 Experimental Setup

Dataset and license compliance. We evaluate on the **Hateful Memes** benchmark [5]. For convenience, we retrieved the official release files via a public Kaggle-hosted mirror.[1] We use the dataset splits as provided: **train** (8500), **dev** (500), and **test** (1000), totaling 10,000 images.

- **Note on Unlabeled Test.** The official test split labels are withheld. Therefore, all quantitative results in this paper are reported on the **dev** split. To mitigate overfitting, we do not perform extensive hyperparameter tuning and report results averaged over 3 random seeds.
- **License note.** We do not redistribute the dataset; we release only code and instructions to reproduce results after obtaining the dataset under its terms.

Table 1: Dataset summary (this release).

Split	#Samples	Modalities	Labels
Train	8500	img + text	provided
Dev	500	img + text	provided
Test	1000	img + text	unlabeled

Embedding-first pipeline. We precompute and cache frozen embeddings for both modalities, storing L2-normalized vectors (fp16). This decouples training from image I/O and heavy encoder execution and enables rapid iteration.

[1] https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset/data

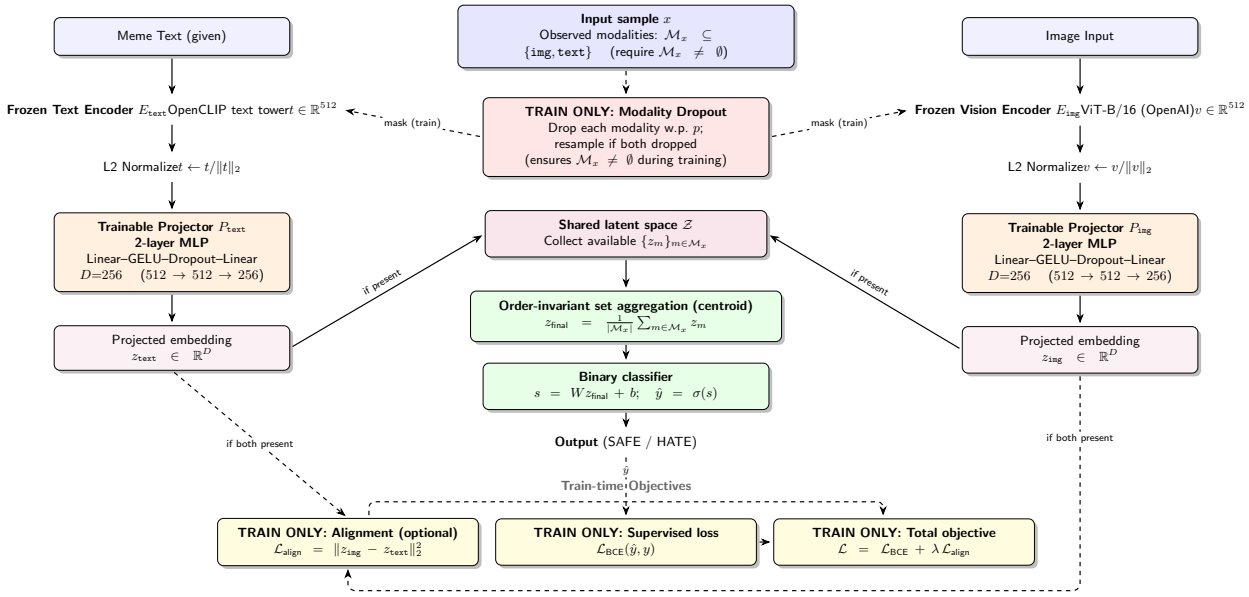

Figure 2: The UniSafe Framework. Frozen encoders ($E_{\texttt{text}}$ OpenCLIP, $E_{\texttt{img}}$ ViT-B/16) extract d=512 normalized embeddings. Trainable 2-layer MLP projectors (Linear–GELU–Dropout–Linear) map each modality into a shared latent space $\mathcal{Z}$ (D=256). UniSafe aggregates the available set via order-invariant centroid averaging, enabling any-subset inference. Training uses modality dropout (p) and an optional alignment objective ($\mathcal{L}_{\text{align}}$) when both modalities are present.

Baselines and fairness. To isolate the effectiveness of our fusion strategy, we compare UniSafe against standard unimodal and multimodal baselines:

- **Unimodal heads:** image-only and text-only heads on frozen embeddings.
- **Late fusion (concat-MLP):** concatenation of image and text embeddings followed by an MLP classifier.

All models use the same cached embeddings, the same train/dev split, and the same optimization budget (epochs/seeds).

Implementation details. We set the latent dimension D=256, hidden width 512, modality dropout p=0.30, and alignment weight λ=0.10. Models are trained for 20 epochs with a batch size of 512. We report mean±std performance over three random seeds $\{0, 1, 2\}$.

Deployment robustness. UniSafe exposes a single operational interface: given any non-empty subset of modalities, it produces a hate probability using one unified checkpoint. Missing modalities are handled by simply omitting their embeddings—no imputation, routing logic, or modality-specific heads are required. This simplifies production resilience (e.g., handling broken image links), as the decision path remains identical across all data availability regimes.

5 Results

Main performance. Table 2 reports dev performance. UniSafe matches late fusion while adding an operational benefit: a *single checkpoint* supports any-subset inference (image-only, text-only, or both). We report Accuracy, AUROC, and Macro-F1.

Missing-modality robustness. We evaluate the *same UniSafe checkpoint* under three inference conditions: (i) full evidence {img + text}, (ii) missing image {text} only, and (iii) missing text {img} only. Missingness is simulated by providing only the available cached embedding(s) at inference; UniSafe applies the same set aggregation and classifier regardless of which modalities are present.

Table 2: Main results on Hateful Memes (dev), mean±std over 3 seeds. UniSafe matches late fusion while enabling any-subset inference.

Method	Acc	AUROC	Macro-F1
Img-Only	**0.593 ± 0.008**	0.6518 ± 0.0003	0.565 ± 0.010
Txt-Only	0.566 ± 0.003	0.6455 ± 0.0002	0.535 ± 0.006
Late Fusion	0.591 ± 0.005	0.6853 ± 0.0016	**0.569 ± 0.004**
UniSafe	0.583 ± 0.003	**0.6865 ± 0.0026**	0.549 ± 0.002

To contextualize robustness, we compare against a practical deployment baseline: **Late Fusion** (concat-MLP) when both modalities are present, and a **Fallback** system that routes to a dedicated img-only or text-only checkpoint when a modality is unavailable. Table 3 reports AUROC means (3 seeds). To avoid repeating the same information, Figure 3 visualizes UniSafe's *gain* over this baseline in each regime.

Discussion. UniSafe achieves late-fusion parity on full inputs while removing the need for routing across multiple checkpoints. Under missing modalities, UniSafe is competitive with the fusion + fallback baseline: it improves text-only performance when the image is unavailable, and is approximately on par for image-only inputs.

Additional Experiments. We add (i) targeted ablations isolating the contribution of *alignment* and *modality dropout*, and (ii) a

Table 3: Any-subset robustness (dev AUROC, mean over 3 seeds). UniSafe uses one checkpoint across full and missing-modality regimes. Baselines use Late Fusion when both modalities are present, and fall back to dedicated unimodal checkpoints when one is missing.

Scenario (AUROC)	Baseline (Fusion/Fallback)	UniSafe (One Model)
Full input (img+text)	0.6853*	**0.6865**
Missing text (img only)	**0.6518**†	0.6509
Missing image (text only)	0.6455‡	**0.6489**

*Late Fusion (concat-MLP). †Dedicated img-only checkpoint. ‡Dedicated text-only checkpoint.

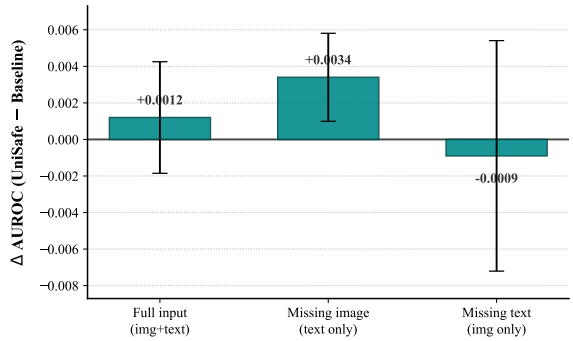

Figure 3: Robustness gains under missing modalities (dev). ΔAUROC of UniSafe relative to the fusion/fallback baseline (UniSafe − Baseline) across three inference regimes. Error bars reflect variability across seeds.

single-checkpoint *Imputed Fusion* baseline that supports missing inputs via zero-imputation + mask bits. We use the identical experimental setup (frozen encoders, 20 epochs, 3 seeds) defined in Section 4.

- **Ablations and single-checkpoint baseline.** Table 4 shows that removing *modality dropout* reduces robustness under missing inputs (especially image-only), while removing *alignment* has negligible impact on this dev split. Imputed Fusion provides a simpler one-checkpoint baseline that is close on full inputs but slightly weaker under missing modalities.
- **Missingness curves.** The all-or-nothing removal in Table 3 is coarse. We therefore simulate *independent* modality failures by dropping each modality with probability r (resampling if both would be absent) and evaluate dev AUROC as a function of $r \in [0, 1]$. Figure 4 shows a smooth degradation as missingness increases. Across the full range of r, UniSafe tracks Imputed Fusion closely and maintains a small but consistent advantage, indicating that the any-subset set-aggregation objective remains stable under increasingly partial evidence.
- **Takeaway.** Across these controlled variants, *modality dropout* is the key ingredient for missing-modality robustness, while alignment (as implemented here) does not materially change dev performance. UniSafe's main advantage remains the *any-subset, one-checkpoint* inference contract with competitive accuracy.

6 Ethics, Limitations, and Broader Impact

This work uses datasets that contain harmful and offensive content. We do not reproduce slurs or hateful imagery in text or figures.

Table 4: Extension experiments (dev AUROC, mean±std over 3 seeds). Ablations isolate the contribution of alignment and modality dropout. Imputed Fusion is a single-checkpoint baseline that supports missing inputs via imputation + mask bits.

Method	Full (img+txt)	Img only	Txt only
UniSafe (p=0.30, λ=0.10)	0.6865 ± 0.0026	0.6509 ± 0.0063	0.6489 ± 0.0024
UniSafe w/o align (p=0.30, λ=0.00)	0.6869 ± 0.0022	0.6511 ± 0.0062	0.6489 ± 0.0024
UniSafe w/o moddrop (p=0.00, λ=0.10)	0.6892 ± 0.0009	0.6424 ± 0.0039	0.6479 ± 0.0009
Imputed Fusion (single checkpoint)	0.6840 ± 0.0007	0.6462 ± 0.0006	0.6445 ± 0.0025

p is modality-dropout prob.; λ weights $\mathcal{L}_{\text{align}}$. **Imputed Fusion** concatenates $[v; t; \mathbb{1}_{\text{img}}; \mathbb{1}_{\text{text}}]$ and trains with modality dropout; at inference, missing modality embeddings are set to zero and the corresponding mask bit is 0.

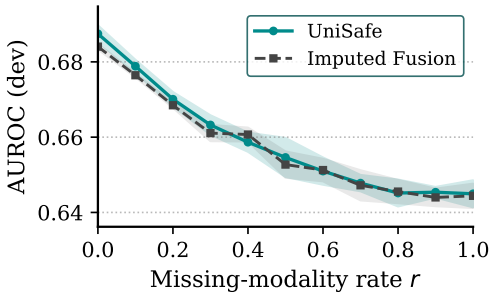

Figure 4: Missingness robustness (dev). AUROC vs. missing-modality rate r for UniSafe and Imputed Fusion (single checkpoint) under independent modality dropout at inference (resampling when both modalities would be missing). Shaded regions indicate variability across seeds.

- **Limitations.** UniSafe relies on frozen encoders: if the foundation representation fails to capture a visual symbol or cultural nuance, the lightweight projectors cannot recover it. Additionally, centroid aggregation is intentionally simple; while it supports any-subset inference, it may underperform in cases where fine-grained *conflict* between text and image is decisive.
- **Broader impact.** UniSafe targets robust moderation under partial inputs; real deployments should include calibration, thresholding, and human-in-the-loop review.

7 Conclusion

We introduced the UniSafe framework, a shared-space projection approach for modality-agnostic hateful content detection. By treating modalities as a set and using centroid aggregation, UniSafe supports robust any-subset inference without fallback routing. On Hateful Memes (dev), UniSafe matches late fusion (AUROC 0.6865 ± 0.0026 vs. 0.6853 ± 0.0016) while using a single checkpoint across full and missing-modality regimes. Ablations show that modality dropout is the primary driver of robustness.

Acknowledgment

This research is supported by NSERC Discovery Grants (RGPIN-2024-04087) and NSERC DND Supplements (DGDND-2024-04087).

References

[1] Meriem Bayoudh, Mariem Knani, Faten Hamdaoui, and Abir Mtibaa. 2022. A Survey on Deep Multimodal Learning for Computer Vision: Advances, Trends, Applications, and Datasets. *The Visual Computer* 38, 8 (2022), 2939–2970. doi:10.1007/s00371-021-02166-7

[2] Rui Cao, Ziqing Fan, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2021. Disentangling Hate in Online Memes. In *Proceedings of the 29th ACM International Conference on Multimedia*. doi:10.1145/3474085.3475625

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal Image-Text Representation Learning. In *European Conference on Computer Vision*. doi:10.1007/978-3-030-58577-8_7

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. arXiv:2212.07143 [cs.CV]

[5] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*. arXiv:2005.04790 [cs.CL] https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html

[6] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557* (2019). https://arxiv.org/abs/1908.03557

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*. arXiv:1908.02265 [cs.CV] https://openreview.net/forum?id=S1eOXNHeUS

[8] Niklas Muennighoff. 2020. Vilio: State-of-the-Art Visio-Linguistic Models Applied to Hateful Memes. *arXiv preprint arXiv:2012.07788* (2020). https://arxiv.org/abs/2012.07788

[9] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2016. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1692–1706. doi:10.1109/TPAMI.2015.2461544

[10] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4439–4455. doi:10.18653/v1/2021.findings-emnlp.379

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. 8748–8763. arXiv:2103.00020 [cs.CV] https://proceedings.mlr.press/v139/radford21a.html

[12] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. doi:10.18653/v1/D19-1514

[13] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. 2017. Deep Sets. In *Advances in Neural Information Processing Systems*. arXiv:1703.06114 [cs.LG] https://papers.nips.cc/paper/6931-deep-sets

A Reproducibility Details

A.1 Cached embedding artifacts

Table 5: Cached embedding artifacts (ViT-B/16, d=512; L2-normalized; fp16).

Artifact	Shape	Dtype
train image embeddings	[8500, 512]	fp16
train text embeddings	[8500, 512]	fp16
dev image embeddings	[500, 512]	fp16
dev text embeddings	[500, 512]	fp16
test image embeddings	[1000, 512]	fp16
test text embeddings	[1000, 512]	fp16

A.2 Compute environment

All experiments were run on a Linux GPU node with:

- **GPU:** NVIDIA Quadro RTX 6000 (24 GB VRAM)
- **Driver / CUDA:** NVIDIA driver 470.256.02; CUDA 11.8
- **Python:** 3.10.19
- **PyTorch:** 2.2.1+cu118

A.3 Training hyperparameters

- **Latent dimension:** D=256, **Hidden width:** 512
- **Optimizer:** AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
- **Learning rate:** 0.001
- **Epochs:** 20 **Batch size:** 512
- **Modality dropout:** p=0.30
- **Alignment weight:** λ=0.10 (optional; see Table 4)
- **Seeds:** 0, 1, 2 **Model selection:** best dev AUROC