Adaptive Trimodal Fusion for Mental-Health Symptom Classification in Memes

Arush Gumber [1 2]

Abstract

Public social-media memes often express anxiety and depression through irony, visual metaphor, and figurative reference, carrying psychological signal that is frequently missed by purely textual analysis. Models for this task may support aggregate public-health research and moderation triage, provided their outputs are treated as uncertain signals rather than individual-level assessments. We study fine-grained anxiety and depression symptom classification on the AxiOM and RESTORE benchmarks and propose **TCAF**, a trimodal classifier that operates on top of three frozen pretrained encoders: a vision encoder over the meme image and a frozen MentalBART text encoder applied separately to OCR text and to figurative reasoning triples. The trainable head combines an Adaptive Gated Trimodal Fusion (AGTF) block that learns a per-dimension soft mask over the concatenated modalities, an additional three-way modality gate, a dual classifier head that blends a linear MLP with learnable class prototypes, and a supervised contrastive auxiliary loss on the fused representation. With only 1.7M trainable parameters, TCAF reaches a test macro-F1 of 0.7067 on AxiOM and 0.6857 on RESTORE, exceeding the prior trimodal state of the art on AxiOM under the same split, with strong results on the official RESTORE test set. We further provide a per-class analysis of modality weighting and a geometric study of the learned class prototypes.

[1]Indraprastha Institute of Information Technology Delhi, India [2]StarkVision Research. Correspondence to: Arush Gumber <arush23136@iiitd.ac.in>.

Accepted at the Workshop on Trustworthy AI for Good (AI4GOOD) at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Public social-media memes often express anxiety and depression through irony, visual metaphor, and figurative reference, and a growing body of prior work studies such content as a potential signal for aggregate analysis of mental-health expression (Mazhar et al., 2025; Yadav et al., 2023). Memes occupy a central place in this discourse. They combine an image, a short caption, and a layer of figurative reference, and they are shared widely enough that automatic recognition of the symptoms they express has both research value, for instance for aggregate public-health research, and operational value, for instance for routing emotionally heavy content to human review on a moderation platform.

The same reach that makes meme content interesting to study also makes it socially sensitive. Because labels such as self-harm and depression are sensitive, errors may have nontrivial downstream consequences if model outputs are surfaced to users or moderators without human review.

Recent work frames the task as fine-grained symptom identification. AxiOM (Mazhar et al., 2025) categorises anxiety memes into six GAD-7 symptoms, while RESTORE (Yadav et al., 2023) provides multilabel PHQ-9 annotations of depression memes. Existing approaches in this space typically combine the OCR text of the meme with structured commonsense reasoning generated by a large language model and then train a moderately large encoder on the resulting input. The current state of the art, M3H (Mazhar et al., 2025), demonstrates the value of GPT-4o reasoning triples for this task. We study a complementary question: whether direct frozen visual features and a small trainable fusion head can improve task-specific training efficiency while retaining the figurative-reasoning signal, in line with prior observations on linear probing of frozen pretrained features (Kumar et al., 2022).

We explore that direction. We propose **TCAF**, a classifier that runs on top of three frozen pretrained encoders. The image is processed by CLIP (Radford et al., 2021), and the OCR and reasoning triples are processed by a frozen MentalBART encoder used by M3H (Mazhar et al., 2025). The trainable head, with 1.7M trainable parameters, has three main components. An *Adaptive Gated Trimodal Fusion*

(AGTF) block applies a per-dimension soft mask to the concatenated modality vectors, allowing the model to suppress noisy coordinates and emphasise informative ones at the level of individual feature dimensions; a small additional three-way modality gate produces a per-sample weighting over the image, OCR, and reasoning streams before AGTF is applied. A *dual classification head* combines a linear MLP with a learnable prototype-based cosine classifier, and the two heads are blended through a learned coefficient. A *supervised contrastive auxiliary objective* (Khosla et al., 2020) on a separate projection of the fused representation directly addresses the near-degenerate centroid geometry observed when frozen pretrained features are used at this dataset scale.

Contributions. We make three contributions:

1. **TCAF**, a 1.7M trainable-parameter trimodal classifier that reaches a test macro-F1 of 0.7067 on AxiOM and 0.6857 on RESTORE under a held-out validation selection protocol, exceeding the prior trimodal state of the art on AxiOM under the same data split and with strong results on the official RESTORE test set.

2. The Adaptive Gated Trimodal Fusion block with an additional three-way modality gate whose softmax outputs admit a per-class reading of which modality received slightly higher weight.

3. Per-class analyses of confusion structure, learned prototype directions, and modality weighting, locating where the model succeeds and where its predictions should be treated with care.

2. Related Work

Memes on social platforms. Memes have become a widely used medium for communication on social media and have been studied along several axes, including detection and categorisation of harmful content, viral spread, and cross-platform behaviour. The Hateful Memes challenge (Kiela et al., 2020) popularised multimodal benchmarks for offensive meme classification, and FigMemes (Liu et al., 2022) provides annotations for six types of figurative language in politically opinionated memes. MemeCLIP (Shah et al., 2024) adapts CLIP representations for meme classification with lightweight feature adapters. Work focused on *mental-health* memes is comparatively recent. RESTORE (Yadav et al., 2023) introduced a fine-grained multilabel benchmark for depression symptoms based on PHQ-9, and AxiOM (Mazhar et al., 2025) introduced a single-label benchmark for anxiety symptoms based on GAD-7. Both datasets use symptom categories derived from clinical questionnaires and were annotated under the protocols described by their authors.

Mental-health expression in online text. A long line of work examines linguistic markers of mental health conditions in social media text. Domain-adapted language models such as MentalBERT (Ji et al., 2022) continue to pretrain BERT (Devlin et al., 2019) on mental-health corpora and provide stronger initializations for downstream symptom classification. Generative variants based on BART (Lewis et al., 2020) have been used for text generation and classification, while M3H (Mazhar et al., 2025) uses a MentalBART (Yang et al., 2024) classifier together with GPT-4o-generated reasoning triples. Our work uses such a frozen text encoder to project OCR and structured reasoning into the same representation space.

Commonsense and figurative reasoning for memes. Decoding the meaning of a meme typically requires both commonsense knowledge and an interpretation of figurative language. COMET (Bosselut et al., 2019) demonstrates that transformers can produce commonsense knowledge-graph inferences when fine-tuned on structured knowledge bases. M3H (Mazhar et al., 2025) extends this idea by prompting GPT-4o to extract three reasoning attributes per meme, namely a cause-effect description, a figurative-understanding gloss, and an inferred mental state, and uses these as additional textual input to its classifier. We use triples generated with the M3H prompt template as a frozen text-feature stream alongside the OCR.

Multimodal fusion. Many approaches to vision-language fusion stack a multi-head attention transformer (Vaswani et al., 2017) or a series of cross-attention layers between modalities. At small dataset sizes, simpler gating schemes are often more robust. The Gated Multimodal Unit of Arevalo et al. (2017) introduces a learned multiplicative gate that controls how each modality contributes to a fused representation. Sparse Mixture-of-Experts routers (Shazeer et al., 2017) are a different approach in which discrete experts are selected by a top-k gate; AGTF is a continuous per-feature gate operating on the concatenated modality vector, not an expert router. Prototype-based classification, popularised by Prototypical Networks (Snell et al., 2017), classifies inputs by similarity to a small set of class anchors. Supervised contrastive learning (Khosla et al., 2020) extends self-supervised contrastive objectives to the supervised case by treating same-class samples as positive pairs, which is particularly useful when class centroids in the feature space are close. The Asymmetric Loss (Ridnik et al., 2021) addresses class imbalance in multilabel classification by applying different focusing strengths to positive and negative samples.

Frozen features with a small trainable head. Kumar et al. (2022) show that frozen pretrained features with a linear probe often outperform full fine-tuning, particularly

when training sets are small and out-of-distribution generalisation is important. Our trimodal head follows this paradigm: the encoders are frozen and only the fusion head, with 1.7M trainable parameters, is updated.

3. Datasets and Task

AxiOM (Mazhar et al., 2025). AxiOM is a single-label dataset of 3,582 anxiety memes, with six fine-grained classes drawn from the GAD-7 questionnaire (Nervousness, Lack of Worry Control, Excessive Worry, Difficulty Relaxing, Restlessness, and Impending Doom). Following the protocol outlined by Mazhar et al. (2025), we use these six categories and the released splits, with 2,148 training, 307 validation, and 614 test memes.

RESTORE (Yadav et al., 2023). RESTORE is a multi-label dataset of depression memes annotated with PHQ-9 symptoms. We use the train and test splits released with the original RESTORE dataset of Yadav et al. (2023) (8,814 train and 662 test memes) and follow Mazhar et al. (2025) in excluding the *Lack of Energy* category, leaving seven symptom categories (Concentration Problem, Eating Disorder, Feeling Down, Lack of Interest, Low Self-Esteem, Self-Harm, Sleeping Disorder). After this exclusion we have 8,692 train and 567 test memes, and we hold out 198 memes from the training portion as a stratified validation split, giving a final 8,494 train / 198 validation / 567 test split.

Inputs. Each meme provides three streams: the image, the OCR text extracted from the image, and a set of GPT-4o figurative reasoning triples generated using the prompt template of Mazhar et al. (2025). Each set of triples covers three attributes: a *cause-effect* description that grounds the meme in a real-world scenario, a *figurative-understanding* gloss that articulates the metaphor or irony at play, and an *inferred mental state* of the implied speaker. We use triples generated with the M3H prompt template and feed them as a textual input alongside the OCR rather than as discrete graph edges, which simplifies the encoding pipeline.

4. Method

Figure 1 summarises TCAF. Each meme produces three frozen feature streams that flow into a small trainable head, which has four parts: per-modality projection and attention pooling, an Adaptive Gated Trimodal Fusion block (with a small additional three-way modality gate for the single-label task), a dual classification head, and an auxiliary contrastive projection.

4.1. Frozen feature extraction

We use three pretrained encoders and never update their parameters. The image is encoded by CLIP ViT-L/14 (Radford et al., 2021), producing a token sequence $\mathbf{X}^{\text{img}} \in \mathbb{R}^{257\times1024}$, namely one CLS token plus a 16×16 patch grid. The OCR text and the GPT-4o reasoning triples are encoded by the same frozen MentalBART encoder (Yang et al., 2024) used by M3H (Mazhar et al., 2025), called twice with different inputs and shared weights, producing $\mathbf{X}^{\text{ocr}} \in \mathbb{R}^{64\times1024}$ and $\mathbf{X}^{\text{tri}} \in \mathbb{R}^{256\times1024}$. Features are computed once per dataset and cached.

4.2. Modality projection and attention pooling

Each token sequence is first projected to a shared $d = 256$ dimensional space through a small block consisting of a linear map, a layer normalisation, and a GELU non-linearity. Each projected sequence is then reduced to a single vector through learned attention pooling,

$$\mathbf{v} \;=\; \sum_{i=1}^{L} \alpha_i\, \mathbf{h}_i, \quad \alpha_i \;=\; \frac{\exp(\mathbf{w}^\top \mathbf{h}_i)}{\sum_j \exp(\mathbf{w}^\top \mathbf{h}_j)}, \tag{1}$$

where $\mathbf{h}_i \in \mathbb{R}^d$ is the projected token at position i, L is the sequence length, and $\mathbf{w} \in \mathbb{R}^d$ is a learned query. Each modality has its own pooling query. The output is three pooled vectors $\mathbf{v}^{\text{img}}, \mathbf{v}^{\text{ocr}}, \mathbf{v}^{\text{tri}} \in \mathbb{R}^d$.

4.3. Adaptive Gated Trimodal Fusion

We concatenate the three pooled vectors into a single trimodal vector

$$\mathbf{c} \;=\; [\mathbf{v}^{\text{img}};\, \mathbf{v}^{\text{ocr}};\, \mathbf{v}^{\text{tri}}] \;\in\; \mathbb{R}^{3d}. \tag{2}$$

A small gating network produces a per-dimension soft mask

$$\mathbf{g} \;=\; \sigma(\mathbf{W}_g\, \mathbf{c} + \mathbf{b}_g) \;\in\; [0,1]^{3d}, \tag{3}$$

where σ is the elementwise sigmoid, $\mathbf{W}_g \in \mathbb{R}^{3d\times3d}$ and $\mathbf{b}_g \in \mathbb{R}^{3d}$ are the learnable gate parameters that are shared across samples and produce a sample-dependent gate vector. Each entry $g_i \in [0,1]$ acts as a continuous relevance score for the i-th coordinate of $\mathbf{c}$: values close to one let the corresponding feature pass through, and values close to zero suppress it. The gated vector

$$\mathbf{x}_g \;=\; \mathbf{g} \odot \mathbf{c} \tag{4}$$

is projected through a small MLP and a final layer normalisation to produce the fused vector

$$\mathbf{f} \;=\; \mathrm{LayerNorm}\big(\mathrm{MLP}(\mathbf{x}_g)\big) \;\in\; \mathbb{R}^d. \tag{5}$$

The gate preserves or attenuates coordinates of the concatenated feature vector before projection: when $g_i \to 1$ the

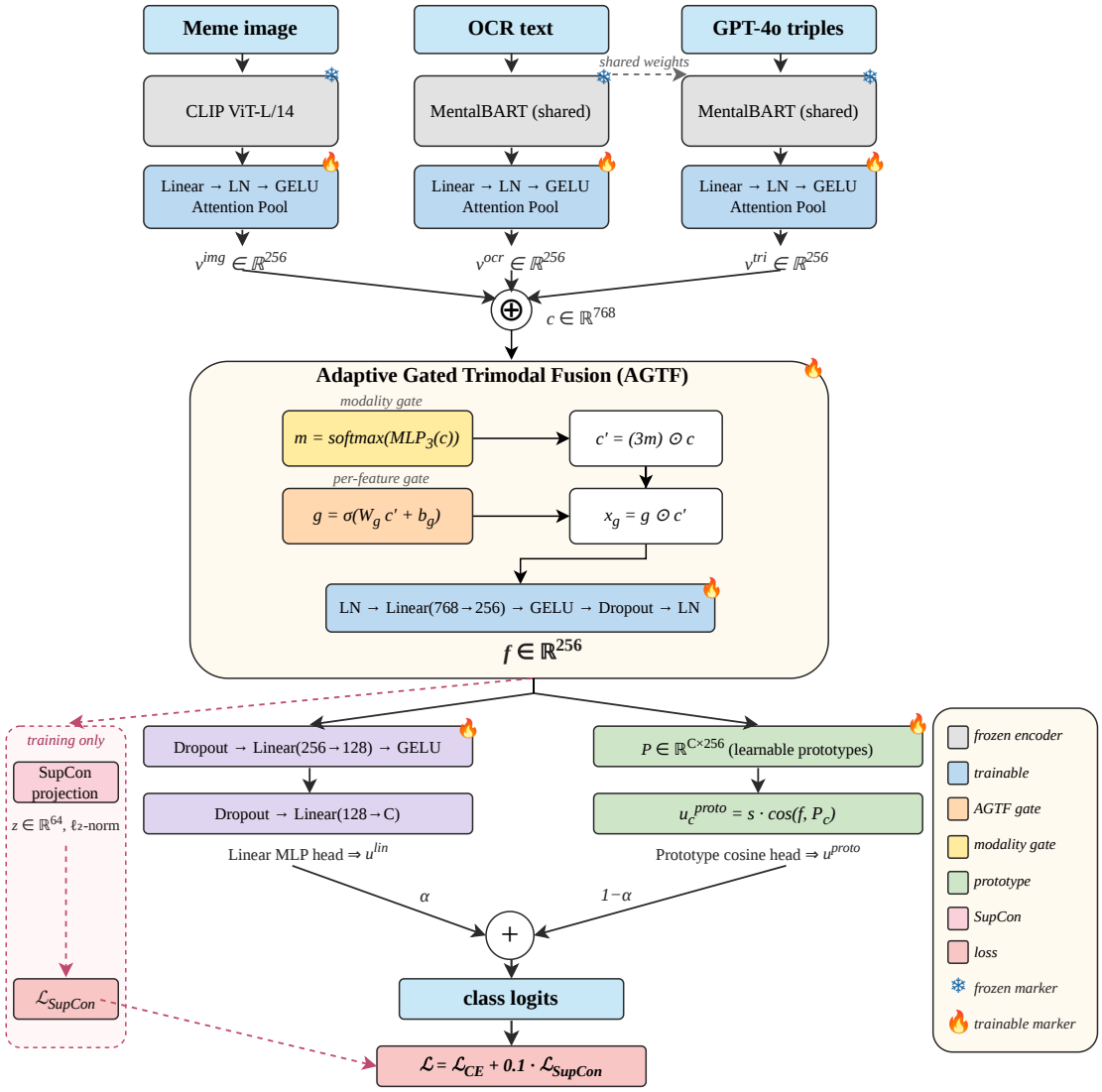


Figure 1. TCAF training-time architecture. Three frozen encoders (the lock marker) feed attention-pooled vectors into the AGTF block, whose output $\mathbf{f}$ is consumed by a dual head combining a linear MLP and a prototype cosine classifier, blended through a learnable coefficient α. The pink branch is active only during training: a separate projection head produces the normalised vector $\mathbf{z}$ used by the supervised contrastive loss.

i-th coordinate of $\mathbf{c}$ passes through, and when $g_i \to 0$ it is suppressed. AGTF is a per-feature relative of the Gated Multimodal Unit (Arevalo et al., 2017) and is distinct from sparse Mixture-of-Experts routers (Shazeer et al., 2017), which select among discrete sub-networks rather than gating individual feature dimensions of a shared vector.

4.4. Three-way modality gate

Section 4.3 defined the base AGTF operation on an input vector $\mathbf{c}$; in the final single-label model, this input is first reweighted by a small three-way modality gate, described below, before the per-feature gate is applied. A two-layer network maps $\mathbf{c}$ to a three-dimensional vector that is normalised through softmax,

$$\mathbf{m} = \mathrm{softmax}\big(\mathrm{MLP}_3(\mathbf{c})\big), \quad \mathbf{m} \in [0,1]^3, \ \sum_{k=1}^{3} m_k = 1, \tag{6}$$

producing a per-sample weighting over the three modalities. Letting $\mathbf{c}_k \in \mathbb{R}^d$ denote the slice of $\mathbf{c}$ that comes from modality $k \in \{\mathrm{img}, \mathrm{ocr}, \mathrm{tri}\}$, we compute

$$\mathbf{c}'_k = (3\, m_k)\, \mathbf{c}_k, \qquad \mathbf{c}' = [\mathbf{c}'_{\mathrm{img}}; \mathbf{c}'_{\mathrm{ocr}}; \mathbf{c}'_{\mathrm{tri}}], \tag{7}$$

and apply the per-dimension AGTF gate to $\mathbf{c}'$ as before, $\mathbf{x}_g = \sigma(\mathbf{W}_g\mathbf{c}' + \mathbf{b}_g) \odot \mathbf{c}'$. The factor 3 keeps the average mass over modalities equal to one, so the modality gate redistributes attention rather than systematically shrinking activations. The entries of $\mathbf{m}$ admit a per-sample reading of which modality received slightly higher weight.

4.5. Dual classification head

For the single-label task we combine a linear MLP head with a trainable prototype cosine head. We maintain a trainable matrix $\mathbf{P} \in \mathbb{R}^{C\times d}$, where C is the number of classes; each row $\mathbf{P}_c \in \mathbb{R}^d$ is a class *prototype* that the model learns to position so that fused representations of class-c memes lie near it on the unit sphere. Prototypes are initialised from $\mathcal{N}(0, 0.02)$ and updated jointly with the rest of the model. The two logit vectors are

$$\mathbf{u}^{\text{lin}} = \text{MLP}_{\text{cls}}(\mathbf{f}), \tag{8}$$

$$u_c^{\text{proto}} = s\,\frac{\mathbf{f}^\top \mathbf{P}_c}{\|\mathbf{f}\|\,\|\mathbf{P}_c\|}, \quad c = 1, \ldots, C, \tag{9}$$

where $s > 0$ is a learnable temperature initialised to 10. Intuitively, $\mathbf{u}^{\text{lin}}$ is a learned linear decision boundary in $\mathbb{R}^d$ while $\mathbf{u}^{\text{proto}}$ is a learned nearest-anchor classifier on the unit sphere; both heads are parametric, the second one's parameters being the prototype directions $\{\mathbf{P}_c\}$. The final logits are a learned blend

$$\hat{\mathbf{y}} = \alpha\,\mathbf{u}^{\text{lin}} + (1-\alpha)\,\mathbf{u}^{\text{proto}}, \qquad \alpha = \sigma(\alpha_0), \tag{10}$$

where $\alpha_0 \in \mathbb{R}$ is a learnable scalar; on AxiOM the model converges to $\alpha \approx 0.61$, leaning slightly toward the linear head while still using the prototype path meaningfully. For the multilabel depression task only the linear head is used, since the prototype head implicitly assumes one active class per sample.

4.6. Auxiliary supervised contrastive loss

For the single-label task the primary objective is weighted cross-entropy with inverse-frequency class weights and a label smoothing of 0.05. We add a supervised contrastive auxiliary loss (Khosla et al., 2020) computed on a separate projection head g_{con} that maps $\mathbf{f}$ through a two-layer MLP with ℓ_2 normalisation to produce $\mathbf{z} \in \mathbb{R}^{64}$. With temperature $\tau = 0.1$, for a batch of N samples and $\mathbf{z}_i = g_{\text{con}}(\mathbf{f}_i)$,

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p/\tau)}{D_i}, \tag{11}$$

$$D_i = \sum_{k\in A(i)}\exp(\mathbf{z}_i^\top \mathbf{z}_k/\tau). \tag{12}$$

where $A(i) = \{1, \ldots, N\} \setminus \{i\}$ and $P(i) \subseteq A(i)$ are the same-class peers of anchor i. The total training loss is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + 0.1\,\mathcal{L}_{\text{SupCon}}. \tag{13}$$

The contrastive term is helpful at this dataset scale because the six anxiety classes have pairwise cosine distances of only 0.0006 to 0.0011 between their centroids in the frozen feature space, so cross-entropy alone provides only a weak gradient signal for the inter-class margin.

For the multilabel depression task, cross-entropy is replaced with the Asymmetric Loss (Ridnik et al., 2021). Let $p_c = \sigma(\hat{y}_c) \in [0, 1]$ be the sigmoid probability for class c, $y_c \in \{0, 1\}$ the binary target, and define a small probability shift $p_{\text{m},c} = \max(p_c - c_{\text{clip}}, 0)$ that zeroes the contribution of very confident negatives. With focusing parameters $\gamma_+, \gamma_- \geq 0$, the per-class loss is

$$\mathcal{L}_{\text{ASL}}^c = \begin{cases} -(1-p_c)^{\gamma_+}\log p_c, & y_c = 1, \\ -\big(p_{\text{m},c}\big)^{\gamma_-}\log(1-p_{\text{m},c}), & y_c = 0. \end{cases} \tag{14}$$

We use $\gamma_+ = 0$, $\gamma_- = 1.5$, and $c_{\text{clip}} = 0.05$, and the total loss is the mean over classes and samples. Per-class sigmoid decision thresholds are tuned on the validation split.

4.7. Training

We train with AdamW at a peak learning rate of 7×10^{-5} for AxiOM and 3×10^{-4} for RESTORE, weight decay 0.05, batch size 32, a cosine learning-rate schedule, and gradient clipping at 1.0. Training stops early when validation macro-F1 has not improved for ten epochs. We select checkpoints using validation macro-F1, and per-class thresholds for RESTORE are also chosen on the validation split.

5. Experiments

5.1. Setup and baselines

We compare TCAF against the strongest baselines reported in Mazhar et al. (2025), which span four input configurations (OCR text alone, OCR with COMET commonsense triples, OCR with LLAVA visual grounding, OCR with GPT-4o reasoning triples) and several backbone families (BERT (Devlin et al., 2019), DeBERTa (He et al., 2021), MentalBERT (Ji et al., 2022), BART (Lewis et al., 2020), MentalBART (Mazhar et al., 2025)). We also compare against Yadav et al. (2023) on RESTORE, the prior state of the art on that benchmark, and against M3H (Mazhar et al., 2025), the prior state of the art on AxiOM. All baseline numbers are taken from Mazhar et al. (2025).

5.2. Main results

Table 1 summarises the comparison. TCAF reaches a test macro-F1 of 0.7067 on AxiOM and 0.6857 on RESTORE,

Table 1. Comparison with prior work on AxiOM and RESTORE. Baseline numbers are taken from Mazhar et al. (2025). TCAF (ours) is the validation-selected single run. ‡M3H is evaluated by Mazhar et al. (2025) on a smaller curated subset of RESTORE (6,863/197/336) that differs from the original Yadav et al. (2023) splits; TCAF uses the official RESTORE train and test splits with the Lack-of-Energy exclusion, so the RESTORE comparison is indicative rather than head-to-head.

Model	**Configuration**	**AxiOM (anxiety)**		**RESTORE (depression)**	
		Macro-F1	W-F1	Macro-F1	W-F1
BERT (Devlin et al., 2019)	OCR only	0.6299	0.6306	0.5816	0.5712
DeBERTa (He et al., 2021)	OCR only	0.6203	0.6186	0.5799	0.5740
MentalBERT (Ji et al., 2022)	OCR only	0.6341	0.6340	0.5914	0.6003
BART (Lewis et al., 2020)	OCR only	0.6282	0.6262	0.4156	0.3835
MentalBART (Mazhar et al., 2025)	OCR only	0.6497	0.6488	0.5419	0.5255
BERT	OCR + GPT-4o	0.6201	0.6202	0.5865	0.5905
MentalBART	OCR + GPT-4o	0.6551	0.6554	0.6326	0.6482
Yadav et al. (2023)	RESTORE benchmark	–	–	0.6358	0.6459
M3H (Mazhar et al., 2025)‡	RAG + fine-tune	0.7000	0.7010	0.6752	0.6879
TCAF (ours)	frozen features + AGTF + dual head	**0.7067**	**0.7072**	**0.6857**	0.6841

Table 2. Incremental component ablation on AxiOM, building up the trainable head one component at a time. The final TCAF configuration has 1.7M trainable parameters.

Configuration	**Macro-F1**
Text-only baseline (OCR alone)	0.5935
+ GPT-4o reasoning triples	0.6250
+ image stream (concat fusion)	0.6670
+ AGTF gate	0.6896
+ supervised contrastive loss	0.6934
+ dual head (linear + prototype)	0.6947
+ label smoothing 0.05	0.6969
+ three-way modality gate	**0.7067**

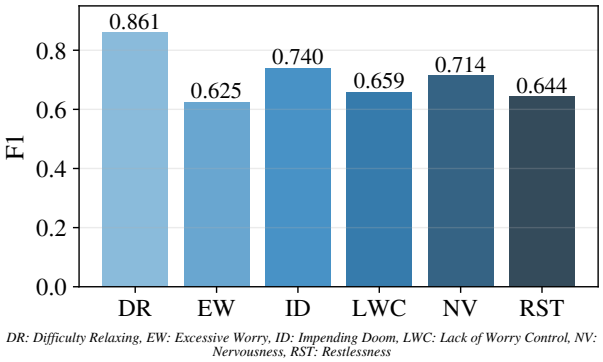


Figure 2. Per-class test macro-F1 on AxiOM.

exceeding M3H by 0.7 absolute points on AxiOM under the same data split. On RESTORE, M3H's reported numbers use a smaller curated subset (see Table 1), so we report TCAF's 0.6857 as a result on the official RESTORE test set rather than a strict head-to-head improvement. The improvement on AxiOM is consistent with the design intuition that frozen features paired with a strong fusion gate generalise better than full fine-tuning at the small training-set scale of this task. On RESTORE, the per-class threshold tuning (see Section 4.7) enables TCAF to handle the multilabel imbalance while still using a single shared head.

5.3. Component ablation

Table 2 ablates the head one component at a time on AxiOM. Going from a text-only baseline to a trimodal concatenation lifts macro-F1 by roughly seven points, replacing concatenation with the AGTF gate adds another two, the contrastive auxiliary adds half a point, and the prototype head plus label smoothing each add a small fraction of a point. The three-way modality gate provides the final lift to 0.7067.

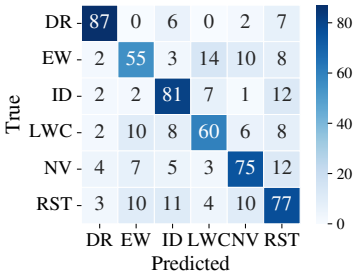
5.4. Per-class behaviour

Figure 2 reports per-class F1 on AxiOM. *Difficulty Relaxing* is by far the easiest class, while *Excessive Worry* and *Restlessness* remain the hardest, reflecting their overlap with neighbouring categories such as *Lack of Worry Control*. Figure 3 shows the corresponding confusion matrix; the same off-diagonal mass appears between Lack of Worry Control and the two harder classes, which is consistent with the observation in Mazhar et al. (2025) that these categories share visual cues such as a person stopping or being interrupted.

6. Analysis

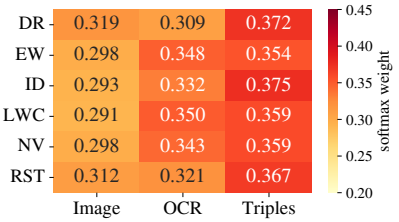
6.1. Per-class modality weighting

Figure 4 shows the per-class average of the modality gate $\mathbf{m}$ on AxiOM, marginalised over test samples of each class. Although the gate is initialised uniformly and is never given modality labels, the trained values reveal small class-specific shifts. Classes whose signal is more visual, such as *Restlessness* and *Difficulty Relaxing*, place slightly larger weight



DR: Difficulty Relaxing, EW: Excessive Worry, ID: Impending Doom, LWC: Lack of Worry Control, NV: Nervousness, RST: Restlessness

Figure 3. Test confusion matrix on AxiOM. Off-diagonal mass concentrates between *Lack of Worry Control* and the two harder classes.



DR: Difficulty Relaxing, EW: Excessive Worry, ID: Impending Doom, LWC: Lack of Worry Control, NV: Nervousness, RST: Restlessness

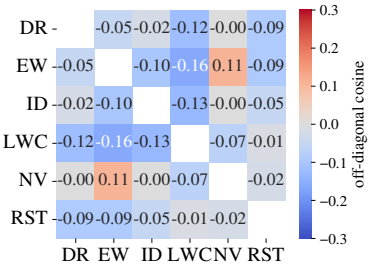
Figure 4. Mean per-class modality gate $\mathbf{m}$ on AxiOM. Differences across classes are small but structured in this run, with image weight slightly higher for visually grounded classes and OCR weight slightly higher for linguistically driven classes.

on the image stream, whereas classes whose signal is more linguistic, such as *Excessive Worry*, place slightly larger weight on OCR. The total deviation from uniformity is small, and the full model benefits from all three streams in our ablations.

6.2. Prototype geometry

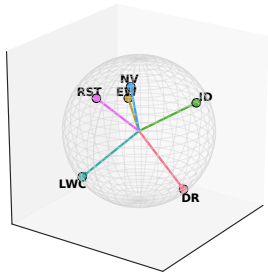
The prototype head computes class scores as the cosine similarity between the normalised fused vector $\mathbf{f}/\|\mathbf{f}\|$ and each normalised row $\mathbf{P}_c/\|\mathbf{P}_c\|$, which means the trained prototypes are most informative when read as directions. Figure 5 reports the pairwise cosine similarity matrix of $\mathbf{P}$ at convergence (off-diagonal entries only), and Figure 6 visualises the same prototypes after a three-component PCA.

After joint optimisation with cross-entropy and the contrastive auxiliary, the six off-diagonal cosines lie in roughly $[-0.16, 0.11]$ with a mean absolute value of about 0.07, so the anchors remain separated rather than collapsing to similar directions. Figure 6 shows the same geometry as a three-dimensional projection: the six directions occupy distinct regions of the projection, which is the geometric counterpart of the small off-diagonal cosines in Figure 5.



DR: Difficulty Relaxing, EW: Excessive Worry, ID: Impending Doom, LWC: Lack of Worry Control, NV: Nervousness, RST: Restlessness

Figure 5. Pairwise cosine similarity of the trained prototype matrix $\mathbf{P}$ on AxiOM. Diagonal entries (equal to one) are masked to highlight the off-diagonal structure that matters for class separation.



DR: Difficulty Relaxing, EW: Excessive Worry, ID: Impending Doom, LWC: Lack of Worry Control, NV: Nervousness, RST: Restlessness

Figure 6. Prototype vectors after PCA projection to three dimensions and renormalisation to unit length. The plot is a visualisation of relative directions, not the space used by the classifier.

7. Discussion

TCAF has a small trainable head and short training time, which makes the task-specific component easier to reproduce and audit. The overall picture from the analyses in Section 6 is that performance is supported by all three modality streams, with small class-specific shifts in modality weight in the run we analyse. The hardest classes are those whose visual and textual signal overlaps with neighbouring categories, and the per-class F1 distribution makes this gap explicit rather than averaging it away.

Why frozen features. At the dataset scales studied here, the trainable head sees only a few thousand labels. In our experiments, updating larger backbones was less stable at this dataset scale; freezing the encoders and learning only the fusion mapping concentrates the trainable surface where the small label budget is best spent, which is consistent with the linear-probing observations of Kumar et al. (2022). This efficiency claim concerns task-specific trainable parameters and training cost; inference still uses the frozen CLIP and MentalBART feature extractors.

Why a per-feature gate. Each modality has different noise behaviour: the image can be a generic stock photograph, the OCR can be a single word, and the reasoning

triples are often more verbose and structured. AGTF lets the model assign each individual feature dimension of the concatenated vector a relevance score, which we view as a finer-grained alternative to a global multiplicative gate. The three-way modality gate preserves the per-feature operation but adds a small modality-level prior.

8. Ethics and Limitations

Our task is fine-grained classification of *symptoms expressed in publicly posted memes*. The labels are dataset annotations of how a meme reads, derived from clinical questionnaires (GAD-7 for AxiOM and PHQ-9 for RESTORE) and produced under the protocols described in the original dataset papers; they are not assessments of any individual's clinical state. The model produced here is a research tool for symptom classification in meme content and is not a diagnostic system.

Mental-health expression is strongly culturally situated. Figurative language, irony, and visual conventions vary across communities, and a model trained on a single annotation pool will systematically misread content from communities that are under-represented in that pool. Both AxiOM and RESTORE are English-language and platform-specific, which limits external validity. Inter-annotator agreement on figurative affective content is also modest, and the classes we predict are not mutually exclusive in human terms.

A model that flags depression or self-harm content carries dual-use risk. We recommend that downstream consumers of any released checkpoint treat its predictions as one of several signals for human review rather than as a stand-alone classifier, and that they couple deployment with calibrated uncertainty estimates and clear non-clinical disclosure. Our main results are from a validation-selected single run; reporting variance over multiple seeds, comparing against zero-shot vision-language model baselines, and testing cross-dataset generalization are important directions we leave to future work.

9. Conclusion

We presented TCAF, a trimodal classifier for fine-grained anxiety and depression symptom classification in memes. With only 1.7M trainable parameters on top of three frozen pretrained encoders, TCAF reaches 0.7067 macro-F1 on AxiOM and 0.6857 macro-F1 on RESTORE, improving on the prior trimodal state of the art on AxiOM under the same split and achieving strong performance on the official RESTORE test set. We provided per-class confusion analyses, a per-class reading of the modality gate, and a geometric study of the trained prototype directions.

References

Arevalo, J., Solorio, T., y Gómez, M. M., and González, F. A. Gated multimodal units for information fusion, 2017. URL https://arxiv.org/abs/1702.01992.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. COMET: Commonsense transformers for automatic knowledge graph construction. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL https://aclanthology.org/P19-1470/.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

He, P., Liu, X., Gao, J., and Chen, W. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. MentalBERT: Publicly available pretrained language models for mental healthcare. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7184–7190, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.778/.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes

challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Liu, C., Geigle, G., Krebs, R., and Gurevych, I. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7069–7086, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.476. URL https://aclanthology.org/2022.emnlp-main.476/.

Mazhar, A., Shaik, Z. H., Srivastava, A., Ruhnke, P., Vaddavalli, L., Katragadda, S. K., Yadav, S., and Akhtar, M. S. Figurative-cum-commonsense knowledge infusion for multimodal mental health meme classification. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 637–648, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714778. URL https://doi.org/10.1145/3696410.3714778.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, October 2021.

Shah, S. B., Shiwakoti, S., Chaudhary, M., and Wang, H. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17320–17332, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.959. URL https://aclanthology.org/2024.emnlp-main.959/.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4080–4090, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yadav, S., Caragea, C., Zhao, C., Kumari, N., Solberg, M., and Sharma, T. Towards identifying fine-grained depression symptoms from memes. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8890–8905, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.495. URL https://aclanthology.org/2023.acl-long.495/.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., and Ananiadou, S. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 4489–4500, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3648137. URL https://doi.org/10.1145/3589334.3648137.

A. Per-class results on RESTORE

Figure 7 shows the per-class F1 of TCAF on RESTORE. Three classes (Concentration Problem, Eating Disorder, and Sleeping Disorder) reach an F1 above 0.80, while the rare *Lack of Interest* class is the hardest, reflecting both its low support and overlap with *Feeling Down*. Figure 8 reports the per-class binary confusion matrices that produced these F1 values; the false-negative rate dominates for the rare classes, which is the expected behaviour of a thresholded sigmoid head under the imbalanced positive rate.

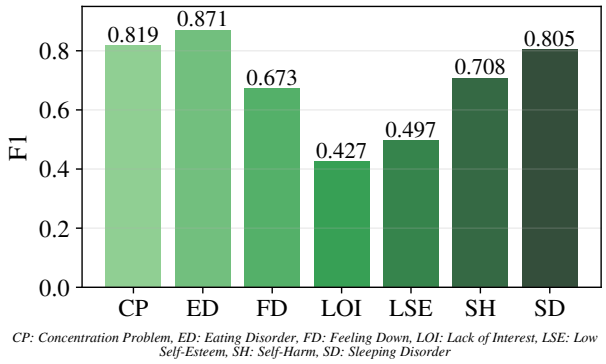


Figure 7. Per-class test macro-F1 on RESTORE. Class abbreviations are listed below the panel.

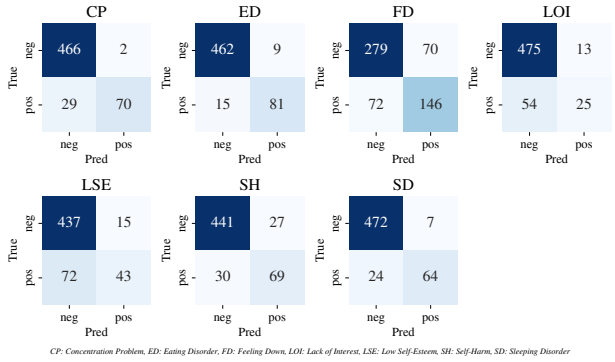


Figure 8. Per-class test confusion matrices on RESTORE. Each panel corresponds to a single binary symptom.

B. Per-class decision thresholds for RESTORE

The per-class sigmoid thresholds for the validation-best configuration are reported in Table 3. They are obtained through a per-epoch search on the validation split during training and frozen for inference.

C. Hyperparameters and reproducibility

For both datasets we use AdamW with weight decay 0.05, gradient clipping at 1.0, and a cosine learning-rate schedule. The peak learning rate is 7×10^{-5} for AxiOM and 3×10^{-4} for RESTORE, with batch size 32. Early stopping triggers after ten epochs without an improvement in validation macro-F1. The supervised contrastive temperature is $\tau = 0.1$ and the auxiliary weight is 0.1. The Asymmetric Loss for RESTORE uses $\gamma_+ = 0$, $\gamma_- = 1.5$, and $c_{\text{clip}} = 0.05$. Weight averaging on AxiOM is computed over the top five checkpoints by validation macro-F1 and accepted only if its validation macro-F1 stays within 0.005 of the best single checkpoint.

Table 3. Per-class sigmoid decision thresholds used at inference on RESTORE.

Class	Threshold
Concentration Problem	0.46
Eating Disorder	0.73
Feeling Down	0.25
Lack of Interest	0.16
Low Self-Esteem	0.34
Self-Harm	0.49
Sleeping Disorder	0.62

D. Trainable parameters versus accuracy

Figure 9 plots trainable parameter count against AxiOM test macro-F1 for TCAF and the strongest fully fine-tuned baselines reported in Mazhar et al. (2025). TCAF occupies the upper-left region: it reaches the highest macro-F1 in the comparison while updating about 70 times fewer trainable parameters than the next strongest model. The plot also makes clear that across the fine-tuned baselines, additional trainable parameters do not translate into proportional accuracy gains at this dataset scale.

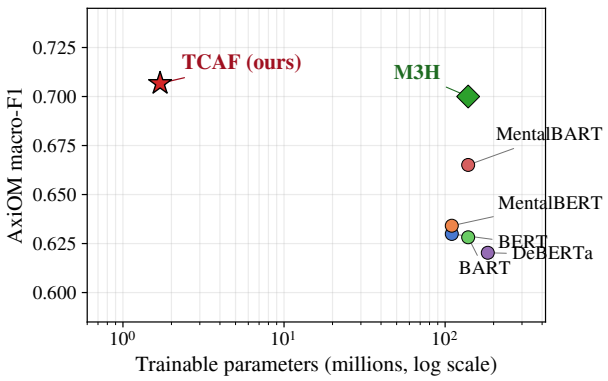
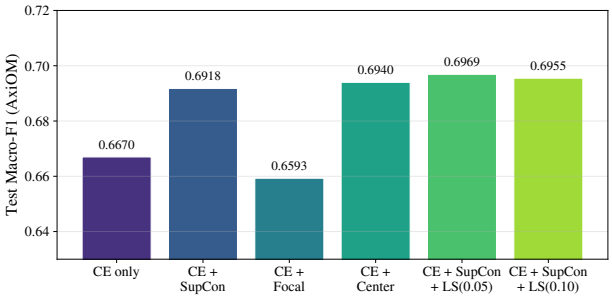
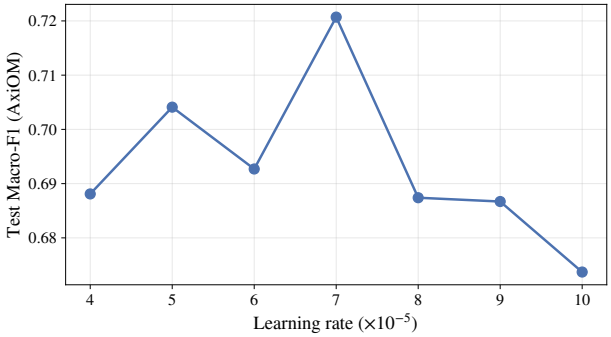


Figure 9. Trainable parameters (millions, log scale) versus AxiOM test macro-F1. TCAF (red star) reaches the highest macro-F1 in the comparison while updating about 70 times fewer trainable parameters than the fully fine-tuned baselines.



(a) Loss function comparison.



(b) Learning-rate sensitivity for TCAF on AxiOM.

Figure 10. Auxiliary studies on AxiOM.

E. Loss function and learning-rate sensitivity

Figure 10 reports two auxiliary studies on AxiOM. Figure 10 (a) compares loss configurations applied to the same TCAF head and the same hyperparameters: weighted cross-entropy with a supervised contrastive auxiliary term gives the most robust combination, while focal loss alone underperforms because the class imbalance in AxiOM is mild (class weights already address it). Figure 10 (b) sweeps the peak learning rate around the chosen value of 7×10^{-5}; performance peaks sharply at this setting and degrades on either side, with rates above 1×10^{-4} destabilising the contrastive signal and rates below 3×10^{-5} underfitting at the early-stopping budget.