TIDC: Test-Time Intervention-based Diagnosis and Calibration for Harmful Meme Detection

Anonymous ACL submission

Abstract

Harmful meme detection requires a stable understanding of how the image and the text work together. Since the text is already given as external input, a natural test is whether predictions remain stable when the corresponding visible text is removed from the image. Motivated by test-time multi-view inference, we build a text-erased reference view by keeping the external text unchanged and removing the visible text. If the model uses the external text in a stable way, this change should not greatly affect its harmfulness prediction. However, we observe clear score shifts and prediction flips, showing cross-view instability under fixed external text. We further find that the cross-view change after neutral probe filtering is related to error risk, so it can serve as a test-time diagnostic signal. Based on this finding, we propose TIDC, a test-time framework that reduces disturbance from reference view construction, selects a sample-specific calibration strength, and predicts from the calibrated original-view score. We also provide theoretical analysis to support the diagnostic signal and calibration design. Experiments on HarM, FHM, and MAMI show that TIDC consistently improves zero-shot performance across frozen multimodal backbones, without parameter updates or long reasoning chains.

1 Introduction

With the rapid growth of social media, memes have become a widely used form of online expression that combines images and text. This makes harmful meme detection an important task in multimodal content understanding and content safety. Unlike standard image-text classification, the harmfulness of a meme is often not determined by a single modality. Instead, it depends on how the image and the text are jointly composed, how they provide context for each other, and what implicit meaning emerges from their interaction (Hee et al., 2024; Yang et al., 2024; Cao et al., 2022). The same visual content may convey different meanings when paired with different text, while the same text may appear harmful or benign under different visual contexts. As shown by the comparison between panels (a) and (b) in Figure 1, changing only the visible text while keeping similar visual content can substantially change the meme semantics and lead to a different harmfulness judgment.

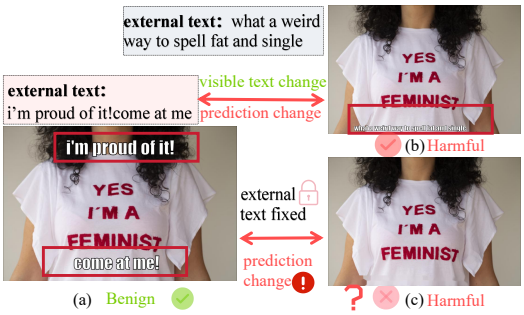

Figure 1: **Motivating examples.** The comparison between (a) and (b) shows that, even when two original memes share similar visual content, differences in textual content can lead to different harmfulness labels. The comparison between (a) and (c) shows that, even when the external text input is fixed, removing the corresponding visible text from the image to construct a reference view can still change the model prediction.

Therefore, the key challenge in harmful meme detection is not only to recognize visual content or understand text, but also to make stable judgments from their interaction.

Recent MLLMs have improved multimodal understanding (Liu et al., 2024; Dai et al., 2023; Chen et al., 2023a), and prior work has advanced harmful meme detection through stronger representations, cross-modal fusion, retrieval, prompting, multi-agent reasoning, and external knowledge (Cao et al., 2022; Yang et al., 2024; Mei et al., 2024; Liu et al., 2025; Garg et al., 2025).

Although these methods are effective, they often improve performance by strengthening the model, augmenting the input, or using more complex rea-

soning procedures. We study a complementary test-time question: when model parameters and external text are fixed, can a controlled image change reveal prediction instability and provide a useful signal for calibrating the original-view score? Inspired by the multi-view idea of test-time augmentation (Shanmugam et al., 2021), we consider whether an additional view of the current meme can expose such instability for a frozen multimodal model. Harmful meme detection provides a natural setting for this question. During inference, the model usually receives both the meme image and external text. Thus, textual content may enter the model through two channels: as visible text in the image and as external text given to the model. Based on this dual-channel setting, we construct a text-erased reference view. Specifically, we keep the external text unchanged, remove detected visible-text regions from the image, and inpaint the removed areas. The original and reference views therefore use the same external text, and mainly differ in whether visible text remains in the image. If the model uses the external text and remaining visual content stably, removing text already reflected in the external input should not cause a large prediction shift.

However, we observe the opposite. As shown by the comparison between panels (a) and (c) in Figure 1, removing the corresponding visible text can change the model prediction even when the external text input remains fixed. This suggests that the model may not stably align the visible text with the external text, and may fail to keep a consistent judgment under fixed external text. We refer to this behavior as cross-view instability. Motivated by these observations, we propose Test-Time Intervention-based Diagnosis and Calibration (TIDC), a training-free framework for harmful meme detection. TIDC first constructs a text-erased reference view under fixed external text, and then compares the model scores between the original view and the reference view. Because the reference view is created by text erasure and local inpainting, the raw cross-view change may include effects caused by reference-view construction. To reduce these effects, TIDC uses a neutral probe to filter the raw cross-view change. The filtered change is then used as a sample-level diagnostic signal to calibrate the original-view score. The final prediction is made from the calibrated original-view score. The whole process is performed only at test time, without updating model parameters or relying on long-form generative reasoning.

We evaluate TIDC on three harmful meme detection benchmarks: HarM, FHM, and MAMI. Experiments show that TIDC improves zero-shot harmful meme detection across frozen multimodal backbones. Further analysis shows that samples with larger filtered cross-view changes are more likely to be misclassified by the original-view prediction, supporting the use of the filtered change as a test-time diagnostic signal.

Our contributions are as follows:

- We show that removing visible text can cause score shifts and prediction flips even when the external text input is fixed, across datasets and model sizes.
- We introduce TIDC, an inference-time calibration method that uses a text-erased reference view to diagnose cross-view instability under fixed external text, reduces reference-view editing effects with a neutral probe, and calibrates the original-view score.
- We provide a theoretical analysis from the view-invariance perspective. We formalize the filtered signal between the original and reference views, and show that the calibration objective can be viewed as a regularized view-invariance loss.

2 Related Work

Harmful meme detection. Prior work has improved harmful meme detection with stronger multimodal representations, cross-modal fusion, retrieval or knowledge augmentation, prompting, and multi-agent reasoning (Kiela et al., 2020; Pramanick et al., 2021; Cao et al., 2022, 2023; Mei et al., 2024; Yang et al., 2024; Huang et al., 2025; Liu et al., 2025). TIDC instead calibrates a frozen model at inference time using the current sample itself.

Test-time inference and calibration. Test-time methods improve predictions using information available at inference time, such as augmented views, adaptation signals, or calibration objectives (Shanmugam et al., 2021; Sun et al., 2020; Wang et al., 2021; Zhang et al., 2022; Shu et al., 2022; Wang et al., 2025; Hekler et al., 2023). Unlike standard view aggregation, TIDC uses the reference view to diagnose instability and calibrate the original-view score.

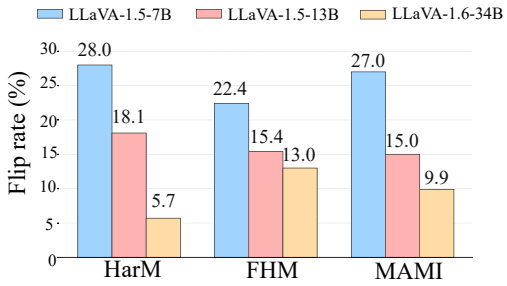

Figure 2: **Prediction flips under fixed external text.** The prediction can flip between the original view and the reference view while the external text input is kept unchanged.

Controlled input edits. Controlled input edits have been used to analyze unstable visual or cross-modal cues (Agarwal et al., 2020; Chen et al., 2023b). TIDC applies this idea by removing detected visible-text regions while keeping the external text fixed, treating the edited image as a diagnostic reference rather than a clean replacement.

3 Method

3.1 Motivation and Overview

TIDC is based on a simple view-invariance principle. When the external text is fixed and already provides the task-relevant textual content, the model's harmfulness judgment should not strongly depend on whether the same text also appears visually in the image. Thus, the original view and the text-erased reference view should produce similar harmfulness scores under the same external text. A large score gap between the two views indicates that the model prediction is sensitive to the view identity, rather than being stable with respect to the underlying meme semantics. In practice, this invariance is often violated. As shown in Figure 2, removing visible text can cause prediction flips even when the external text input is unchanged. Such flips do not determine which view is correct, but they directly reveal cross-view instability under fixed external text.

This motivates using the cross-view score gap as a sample-level diagnostic signal. However, the reference view is constructed by text erasure and local inpainting, so the raw gap may also contain view-specific nuisance introduced by the image editing process. TIDC therefore uses a neutral probe to estimate this construction-induced gap. By comparing the same two views under a fixed neutral text, TIDC filters the raw cross-view gap and obtains a filtered change that better reflects the model's violation of view invariance.

The full TIDC pipeline is shown in Figure 3. TIDC first constructs a text-erased reference view, then compares the original and reference views under both the sample external text and a neutral probe to obtain the filtered change. This signal is used to select a calibration strength and to predict from the calibrated original-view score. The method is applied only at inference time, without parameter updates or long-form generation. We also provide a theoretical analysis of the diagnostic signal and the calibration objective.

3.2 Task Setting and Harmfulness Score

Let $(V_i^{(0)}, T_i, Y_i)$ be the i-th test sample, where $V_i^{(0)}$ is the original meme image, T_i is the external text, and $Y_i \in \{0, 1\}$ is the label, with 1 indicating harmful and 0 benign. The model parameters θ are kept frozen. TIDC uses a continuous harmfulness margin rather than only a discrete predicted label. The continuous score captures how the model score moves across different views, and is therefore more suitable for cross-view diagnosis and calibration.

Let $\ell_\theta(w \mid V, T)$ denote the logit of answer token w under the image-text input (V, T). Let W_H and W_B denote the candidate answer-token sets for the harmful and benign classes. For any candidate set W, we define the aggregated logit as

$$L_\theta(W \mid V, T) = \log \sum_{w \in W} \exp \ell_\theta(w \mid V, T). \quad (1)$$

The harmfulness score is defined as

$$s_\theta(V, T) = L_\theta(W_H \mid V, T) - L_\theta(W_B \mid V, T). \quad (2)$$

When $s_\theta(V, T) > 0$, the model prefers the harmful class. When $s_\theta(V, T) < 0$, the model prefers the benign class. This score can be viewed as a signed margin: its sign gives the preferred class, and its absolute value gives the size of the logit margin.

3.3 Reference View Construction

As shown in Figure 3(A), TIDC constructs a text-erased reference view to examine the model's sensitivity to visible-text presentation under fixed external text. The goal is not to remove all content in the image, but only to process the visible-text regions detected by OCR.

Given the original image $V_i^{(0)}$, TIDC first uses an OCR text detector to locate visible-text regions

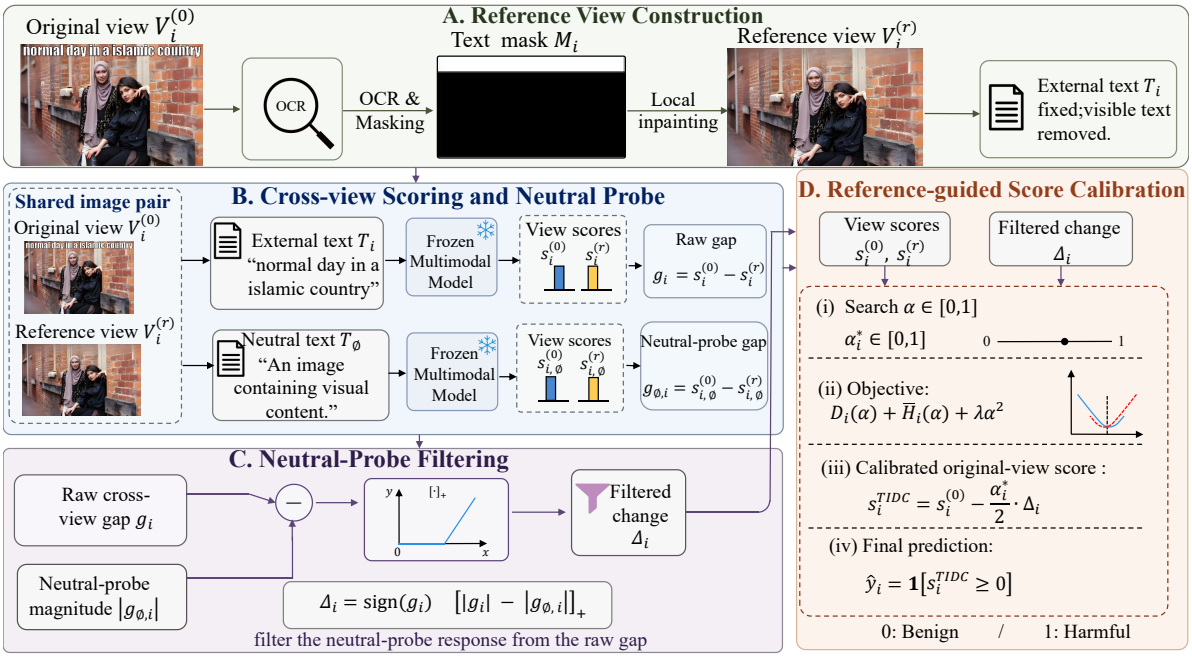

Figure 3: **Overview of TIDC.** Given an original meme image and its external text, TIDC first constructs a text-erased reference view by removing detected visible-text regions while keeping the external text fixed. It then scores the original and reference views under both the sample external text and a fixed neutral probe, obtaining the raw cross-view gap g_i and the neutral-probe gap $g_{\emptyset,i}$. The neutral-probe gap is used to filter construction effects from the raw gap, producing the filtered change Δ_i. Finally, TIDC selects a calibration strength along a symmetric path and predicts from the calibrated original-view score.

in the image. It then moderately expands the detected boxes to cover text boundaries and residual text pixels, and merges all expanded boxes into a text mask M_i. Finally, TIDC performs text erasure and local inpainting only inside the mask, obtaining the reference view:

$$V_i^{(r)} = \Phi(V_i^{(0)}, M_i). \tag{3}$$

Here, Φ denotes the local text-removal operator. It keeps pixels outside the mask unchanged, and fills the removed regions inside the mask with an inpainting model. The original view $V_i^{(0)}$ and the reference view $V_i^{(r)}$ are both paired with the same external text T_i. Thus, this intervention changes only the visible-text regions while keeping the external text received by the model unchanged.

Since the reference view is produced by automatic detection, mask expansion, and inpainting, it may contain local editing artifacts. For example, OCR may miss text or falsely detect non-text regions, the mask may cover nearby regions, and inpainting may change local texture or saliency. For this reason, TIDC uses the reference view as a diagnostic view rather than as a noise-free replacement for the original image.

3.4 Cross-view Signal and Neutral Probe

After obtaining the reference view, TIDC computes harmfulness scores for four image-text pairs for each test sample, as shown in Figure 3(B) and Figure 3(C). All scores are computed by the margin score defined above. Under the sample external text T_i, we compute the original-view score, the reference-view score, and their raw cross-view gap as

$$\begin{aligned} s_i^{(0)} &= s_\theta(V_i^{(0)}, T_i), \\ s_i^{(r)} &= s_\theta(V_i^{(r)}, T_i), \\ g_i &= s_i^{(0)} - s_i^{(r)}. \end{aligned} \tag{4}$$

The sign of g_i gives the direction of the score change, and $|g_i|$ gives its magnitude. A prediction flip is the discrete case where this gap crosses the decision boundary, while the continuous gap also keeps information from samples whose scores move clearly without a label flip.

Because the reference view is produced by text erasure and inpainting, g_i may include responses caused by local editing. To provide a reference for this part, TIDC replaces the sample external text with a fixed neutral text $T_\emptyset$, such as "An image containing visual content.", and evaluates the same

two views again. The corresponding neutral-probe scores and gap are

$$\begin{aligned} s_{i,\emptyset}^{(0)} &= s_\theta(V_i^{(0)}, T_\emptyset), \\ s_{i,\emptyset}^{(r)} &= s_\theta(V_i^{(r)}, T_\emptyset), \\ g_{\emptyset,i} &= s_{i,\emptyset}^{(0)} - s_{i,\emptyset}^{(r)}. \end{aligned} \tag{5}$$

The gap $g_{\emptyset,i}$ measures how much score difference still appears between the same original and reference views when the sample-specific external text is replaced by a fixed neutral text.

TIDC uses it to filter the raw gap. Let $[x]_+ = \max(x, 0)$ and set $\mathrm{sign}(0) = 0$. The filtered change is defined as

$$\Delta_i = \mathrm{sign}(g_i)\left[|g_i| - |g_{\emptyset,i}|\right]_+. \tag{6}$$

This definition keeps the direction of the raw gap and subtracts the magnitude of the neutral-probe response. When $|g_{\emptyset,i}| \geq |g_i|$, we have $\Delta_i = 0$. When $g_{\emptyset,i} = 0$, we have $\Delta_i = g_i$. Therefore, Δ_i is the sample-level cross-view signal used for calibration.

3.5 Reference-guided Score Calibration

TIDC uses the filtered change Δ_i to calibrate the original-view score. The calibration strength $\alpha \in [0, 1]$ controls how much of the filtered change is used for the current sample: $\alpha = 0$ keeps the scores unchanged, while $\alpha = 1$ removes the filtered component of the original-reference gap along the symmetric calibration path. Since Δ_i is already obtained after subtracting the neutral-probe response from the raw gap, we restrict α to $[0, 1]$ to avoid reversing the correction direction or using a correction larger than the diagnosed signal.

For a candidate α, TIDC defines

$$\begin{aligned} \tilde{s}_i^{(0)}(\alpha) &= s_i^{(0)} - \frac{\alpha}{2}\Delta_i, \\ \tilde{s}_i^{(r)}(\alpha) &= s_i^{(r)} + \frac{\alpha}{2}\Delta_i. \end{aligned} \tag{7}$$

This path changes the original-reference gap from g_i to $g_i - \alpha\Delta_i$, while preserving the score center of the two views, where $\alpha \in [0, 1]$ controls the correction strength. Given α, each calibrated score is converted to a harmfulness probability

$$\tilde{p}_i^{(z)}(\alpha) = \sigma(\tilde{s}_i^{(z)}(\alpha)), \qquad z \in \{0, r\}.$$

Let $Q_i^{(z)}(\alpha) = \mathrm{Bern}(\tilde{p}_i^{(z)}(\alpha))$ be the corresponding Bernoulli predictive distribution. We denote by $D_i(\alpha)$ the Jensen-Shannon divergence (JSD) between $Q_i^{(0)}(\alpha)$ and $Q_i^{(r)}(\alpha)$, and by $\bar{H}_i(\alpha)$ the average entropy of the two Bernoulli distributions.

The sample-specific calibration strength is selected by

$$\alpha_i^\star = \arg\min_{\alpha \in \mathcal{A}} \left[D_i(\alpha) + \bar{H}_i(\alpha) + \lambda\alpha^2\right]. \tag{8}$$

Here, $D_i(\alpha)$ measures the remaining disagreement between the two views, $\bar{H}_i(\alpha)$ controls prediction uncertainty, and the quadratic term penalizes large corrections.

After selecting $\alpha_i^\star$, TIDC predicts from the calibrated original-view score:

$$s_i^{\mathrm{TIDC}} = s_i^{(0)} - \frac{\alpha_i^\star}{2}\Delta_i, \qquad \hat{y}_i = \mathbf{1}[s_i^{\mathrm{TIDC}} \geq 0]. \tag{9}$$

3.6 Theoretical Analysis

We analyze TIDC from the view-invariance perspective. We use the notation defined above. Full proofs are given in Appendix A. For analysis, write the original view as $V_i^{(0)} = (X_i, O_i)$ and the ideal reference view as $V_i^{(r)} = (X_i, \emptyset)$.

Assumption 1. We use the following idealized condition: if T_i already contains the task-relevant text, then the visible-text region should not provide additional label information:

$$\begin{aligned} P(Y_i \mid X_i, O = o, T_i) = P(Y_i \mid X_i, T_i), \\ o \in \{O_i, \emptyset\}. \end{aligned} \tag{10}$$

This assumption concerns label information, not the model's response to the visual form of text.

Proposition 1. Under Assumption 1,

$$\begin{aligned} P(Y_i \mid V_i^{(0)}, T_i) &= P(Y_i \mid V_i^{(r)}, T_i) \\ &= P(Y_i \mid X_i, T_i). \end{aligned} \tag{11}$$

Thus, if $Z \in \{0, r\}$ denotes the view identity, then $Y_i \perp Z \mid X_i, T_i$ and $I(Y_i; Z \mid X_i, T_i) = 0$.

Assumption 2. The constructed reference view may introduce local editing effects. We write

$$g_i = \tau_i + \epsilon_i, \tag{12}$$

where τ_i is the view-dependent response and ϵ_i is the construction nuisance. The neutral probe bounds this nuisance:

$$|\epsilon_i| \leq |g_{\emptyset,i}|. \tag{13}$$

This assumption formalizes the neutral-probe gap as a sample-level upper estimate of construction-induced nuisance. It is an analytical condition rather than a claim that the bound always holds exactly.

Proposition 2. Let $b_i = |g_{\emptyset,i}|$. Then

$$|\Delta_i| = \min_{e \in [-b_i, b_i]} |g_i - e|. \tag{14}$$

When $|g_i| > b_i$, Δ_i has the same sign as τ_i and $|\Delta_i| \leq |\tau_i|$. The calibration path also satisfies

$$\tilde{s}_i^{(0)}(\alpha) - \tilde{s}_i^{(r)}(\alpha) = g_i - \alpha\Delta_i, \tag{15}$$

while preserving the two-view score center.

Proposition 3. For a view-invariant predictor R_i, define

$$\begin{aligned}\mathcal{L}_i(\alpha, R_i) = \frac{1}{2}\mathrm{CE}\big(Q_i^{(0)}(\alpha), R_i\big) \\ + \frac{1}{2}\mathrm{CE}\big(Q_i^{(r)}(\alpha), R_i\big).\end{aligned} \tag{16}$$

Then

$$\min_{R_i} \mathcal{L}_i(\alpha, R_i) = D_i(\alpha) + \bar{H}_i(\alpha). \tag{17}$$

Therefore, TIDC selects α by minimizing

$$\alpha_i^\star = \arg\min_{\alpha \in \mathcal{A}} \left[\min_{R_i} \mathcal{L}_i(\alpha, R_i) + \lambda\alpha^2 \right], \tag{18}$$

which is a regularized view-invariant log-loss objective.

4 Experiment

4.1 Experimental Setup

We evaluate TIDC on three harmful meme detection benchmarks: HarM (Pramanick et al., 2021), FHM (Kiela et al., 2020), and MAMI (Fersini et al., 2022). Following MIND (Liu et al., 2025), we use the same data splits and report Accuracy and macro-averaged F1.

All experiments are conducted in the zero-shot setting with frozen model parameters: we do not fine-tune on training samples or tune method settings with test labels. Given the image and external text, each model predicts with a fixed classification instruction, and the classification score is computed using the harmfulness score in Section 3.2.

We apply TIDC to several frozen multimodal backbones, including LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-1.6-34B (Liu et al., 2024), and Qwen2.5-VL-7B (Bai et al., 2025). We also compare with commercial models, open-source multimodal models, and the zero-shot harmful meme detection method MIND (Liu et al., 2025). For each test sample, TIDC computes scores for the original view, the text-erased reference view, and the corresponding neutral-probe inputs. It then performs sample-level calibration at inference time.

TIDC implementation. We construct the reference view with a fixed OCR-based text-erasure pipeline. CRAFT (Baek et al., 2019) detects visible-text regions, whose boxes are moderately expanded and merged into a binary mask. LaMa (Suvorov et al., 2022) then inpaints only the masked pixels, leaving other pixels unchanged. The same text detection, mask expansion, and inpainting settings are used across all datasets and backbones. For calibration, we search $\mathcal{A} = \{0, 0.1, \ldots, 1.0\}$, choose the smallest α in case of ties, and set $\lambda = 0.1$ for all experiments. Details are in Appendix B, with code at `https://anonymous.4open.science/r/TIDC_Arr/`.

4.2 Main Results

Table 1 reports the main zero-shot harmful meme detection results. Overall, TIDC improves macro-F1 for the corresponding frozen LLaVA backbones, and improves or matches Accuracy. The gains are especially clear for smaller open-source models. This shows that test-time cross-view diagnosis and sample-level calibration can improve frozen multimodal models without parameter updates or long-form reasoning.

TIDC performs strongly on HarM. After calibration, the LLaVA backbones clearly outperform their vanilla counterparts, and reach or exceed several stronger baselines in the table. This suggests that model scale is not the only factor in zero-shot harmful meme detection. Even with fixed model parameters, correcting sample-level instability at test time can improve prediction reliability.

On FHM and MAMI, TIDC also improves the open-source backbones. Commercial models remain strong on some metrics, but TIDC reduces the gap between frozen open-source models and stronger systems. These results show that TIDC can serve as a lightweight test-time calibration module for harmful meme detection.

Model	HarM		FHM		MAMI	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
GPT-4o (OpenAI, 2023)	67.51	60.29	68.80	68.25	81.00	81.00
Gemini-1.5-Flash (Reid et al., 2024)	66.10	64.18	60.20	58.90	76.40	74.29
LLaVA-1.5-7B (Liu et al., 2024)	59.23	49.44	53.80	45.51	52.90	41.53
InstructBLIP-7B (Dai et al., 2023)	51.13	50.99	52.00	48.85	53.10	46.93
MiniGPT-v2-7B (Chen et al., 2023a)	60.12	52.39	51.30	47.88	57.40	52.22
OpenFlamingo-9B (Awadalla et al., 2023)	63.42	54.36	50.50	49.52	54.70	49.88
LLaVA-1.5-13B (Liu et al., 2024)	62.28	50.45	55.20	53.01	60.10	55.52
InstructBLIP-13B (Dai et al., 2023)	64.92	49.61	55.40	51.89	60.00	57.97
MIND (LLaVA-1.5-13B) (Liu et al., 2025)	68.93	65.19	60.80	60.71	68.90	68.84
LLaVA-1.6-34B (Liu et al., 2024)	67.51	61.59	**64.00**	63.51	71.30	71.28
LLaVA-1.5-7B + TIDC	71.19	68.79	58.80	58.35	63.40	62.34
LLaVA-1.5-13B + TIDC	**71.75**	**70.15**	61.40	61.39	68.70	68.69
LLaVA-1.6-34B + TIDC	68.64	68.22	**64.00**	**63.99**	**73.90**	**73.80**

Table 1: **Zero-shot harmful meme detection results.** We report Accuracy and macro-averaged F1. Best overall results are underlined, and best results among the listed open-source entries are in **bold**.

Backbone	Setting	HarM		FHM		MAMI	
		Acc.	F1	Acc.	F1	Acc.	F1
LLaVA-1.5-7B	Vanilla	59.23	49.44	53.80	45.51	52.90	41.53
	TIDC	**71.19**(+11.96)	**68.79**(+19.35)	**58.80**(+5.00)	**58.35**(+12.84)	**63.40**(+10.50)	**62.34**(+20.81)
LLaVA-1.5-13B	Vanilla	62.28	50.45	55.20	53.01	60.10	55.52
	TIDC	**71.75**(+9.47)	**70.15**(+19.70)	**61.40**(+6.20)	**61.39**(+8.38)	**68.70**(+8.60)	**68.69**(+13.17)
LLaVA-1.6-34B	Vanilla	67.51	61.59	64.00	63.51	71.30	71.28
	TIDC	**68.64**(+1.13)	**68.22**(+6.63)	**64.00**(+0.00)	**63.99**(+0.48)	**73.90**(+2.60)	**73.80**(+2.52)
Qwen2.5-VL-7B	Vanilla	68.64	64.65	66.20	66.20	67.30	66.38
	TIDC	**70.90**(+2.26)	**69.95**(+5.30)	**66.80**(+0.60)	**66.78**(+0.58)	**71.30**(+4.00)	**71.24**(+4.86)

Table 2: **Scalability analysis of TIDC.** We report Accuracy and macro-averaged F1 across model scales and architectures. Green numbers indicate absolute gains over the vanilla baseline.

4.3 Scalability Across Backbones

Table 2 shows the effect of TIDC across model scales and architectures. TIDC improves macro-F1 in all settings, and improves or matches Accuracy. This indicates that the filtered cross-view signal is not tied to one specific backbone.

Smaller LLaVA backbones obtain larger gains. This suggests that weaker models are more affected by cross-view instability, and benefit more from sample-level calibration. The larger LLaVA-1.6-34B backbone still gains from TIDC, especially in F1. Thus, increasing model scale can reduce some errors, but does not fully remove cross-view instability under fixed external text.

TIDC also improves Qwen2.5-VL-7B. This shows that the method is not limited to the LLaVA family. Since TIDC only uses score changes from the frozen model, it can be used as a lightweight test-time module for different multimodal backbones. Overall, TIDC complements backbone scaling: scaling improves the base model, while TIDC improves sample-level prediction stability.

4.4 Analysis

Figure 4 shows how prediction errors change with the filtered cross-view signal. Within each dataset, we sort test samples by the pre-calibration value of $|\Delta|$ and divide them into five equal-size groups. The vanilla error rate increases with $|\Delta|$, indicating that the filtered cross-view change is a useful diagnostic signal for original-view instability. Figure 4 also shows the effect of TIDC on the same groups. TIDC gives lower error rates than vanilla across the full range of $|\Delta|$. The reduction is more visible in groups with larger $|\Delta|$. This matches the design of TIDC: samples with larger filtered cross-view discrepancy receive stronger sample-level correction. The faint lines show that the trend also appears within each dataset.

Method	HarM		FHM		MAMI	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Vanilla	59.23	49.44	53.80	45.51	52.90	41.53
Raw-gap calibration	65.54 (+6.31)	64.45 (+15.01)	57.90 (+4.10)	57.12 (+11.61)	62.50 (+9.60)	61.45 (+19.92)
Fixed-strength calibration ($\alpha = 1$)	69.49 (+10.26)	66.10 (+16.66)	58.10 (+4.30)	57.84 (+12.33)	61.50 (+8.60)	59.63 (+18.10)
TIDC	**71.19** (+11.96)	**68.79** (+19.35)	**58.80** (+5.00)	**58.35** (+12.84)	**63.40** (+10.50)	**62.34** (+20.81)

Table 3: **Ablation study on LLaVA-1.5-7B.** We report Accuracy and macro-averaged F1. Raw-gap calibration removes neutral-probe filtering, and fixed-strength calibration fixes $\alpha = 1$. Numbers in green indicate absolute gains over the vanilla baseline.

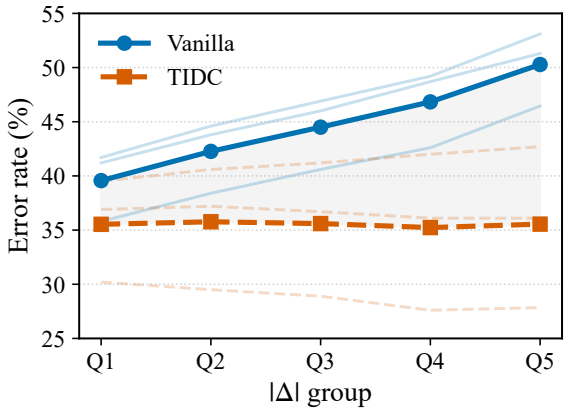

Figure 4: **Diagnostic and mitigation effects of the filtered cross-view signal.** Results are computed with LLaVA-1.5-7B on the full test split. Samples are sorted by the pre-calibration value of $|\Delta|$ and split into five equal-size groups within each dataset. Bold lines show the average over HarM, FHM, and MAMI, and faint lines show per-dataset trends.

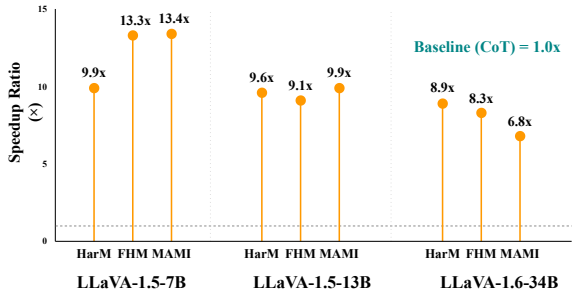

Figure 5: **Inference efficiency.** We report the relative speedup of TIDC over CoT-style prompting. The CoT latency is normalized to $1.0\times$. Higher values indicate faster inference. TIDC avoids long-form autoregressive reasoning and only performs direct scoring over candidate answers.

4.5 Ablation Study

Table 3 reports the ablation results on LLaVA-1.5-7B. Raw-gap calibration removes neutral-probe filtering and directly uses the raw cross-view gap g_i. Fixed-strength calibration keeps the filtered change but fixes α to 1.

Both variants outperform the vanilla baseline, showing that the reference-view gap is not merely noise. Raw-gap calibration is lower than TIDC, which indicates that the unfiltered gap is not the best calibration signal. Fixed-strength calibration is also lower than TIDC, suggesting that using a single correction strength for all samples loses part of the gain. Full TIDC obtains the best results on all three datasets.

4.6 Inference Efficiency

Figure 5 compares the inference efficiency of TIDC and CoT-style prompting. The CoT latency is normalized to $1.0\times$, so a higher value means that TIDC is faster. TIDC achieves clear speedups across datasets and backbones.

The efficiency comes from the inference form of TIDC. TIDC computes scores for the original view, the reference view, and the neutral-probe inputs. Each pass is a direct scoring query over candidate answers. It does not generate a long reasoning chain. CoT-style prompting instead relies on autoregressive generation, which is the main source of latency. Thus, TIDC can be used as a low-overhead test-time calibration module for frozen multimodal models.

5 Conclusion

We examined harmful meme detection under fixed external text and found that erasing visible meme text exposes cross-view instability in frozen multimodal models. We proposed TIDC, a test-time framework that builds a text-erased reference view, filters raw cross-view changes with a neutral probe, and calibrates the original-view score. Experiments on HarM, FHM, and MAMI show that TIDC improves zero-shot performance across frozen backbones without parameter updates or long-form reasoning. Future work will study stronger reference-view construction and extend this diagnostic idea to broader multimodal safety tasks.

Limitations

TIDC uses an OCR-based text-erasure pipeline to construct the reference view. While this setting already provides a useful test-time diagnostic signal, future work could further improve reference-view construction with stronger OCR, text segmentation, and inpainting methods, especially for stylized, small, or multilingual meme text. In addition, our experiments focus on standard harmful meme detection benchmarks, and extending the framework to broader multimodal safety scenarios is a promising direction.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9687–9695. Computer Vision Foundation / IEEE.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9365–9374. Computer Vision Foundation / IEEE.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5244–5252. ACM.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 321–332. Association for Computational Linguistics.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *CoRR*, abs/2310.09478.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023b. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 627–638. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 533–549. Association for Computational Linguistics.

Rahul Garg, Trilok Padhi, Hemang Jain, Ugur Kursuncu, and Ponnurangam Kumaraguru. 2025. Just kiddin' : Knowledge infusion and distillation for detection of indecent memes. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 23067–23086. Association for Computational Linguistics.

Ming Shan Hee, Rui Cao, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Understanding (dark) humour with internet meme analysis. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 1276–1279. ACM.

Achim Hekler, Titus J. Brinker, and Florian Buettner. 2023. Test time augmentation meets post-hoc calibration: Uncertainty quantification under real-world conditions. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14856–14864. AAAI Press.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2025. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 7321–7330. Association for Computational Linguistics.

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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. MIND: A multi-agent framework for zero-shot harmful meme detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 923–947. Association for Computational Linguistics.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5333–5347. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Findings of ACL, pages 2783–2796. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Divya Shanmugam, Davis W. Blalock, Guha Balakrishnan, and John V. Guttag. 2021. Better aggregation in test-time augmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1194–1203. IEEE.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 9229–9248. PMLR.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3172–3182. IEEE.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. 2025. TAPT: test-time adversarial prompt tuning for robust inference in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19910–19920. Computer Vision Foundation / IEEE.

Chuanpeng Yang, Yaxin Liu, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2024. Uncertainty-guided modal rebalance for hateful memes detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4361–4371. Association for Computational Linguistics.

Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. MEMO: test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

A Theoretical Proofs

This appendix proves the propositions in Section 3.6. The notation follows the main text.

A.1 Proof of Proposition 1

Recall that

$$V_i^{(0)} = (X_i, O_i), \qquad V_i^{(r)} = (X_i, \emptyset). \tag{19}$$

By Assumption 1,

$$P(Y_i \mid X_i, O = o, T_i) = P(Y_i \mid X_i, T_i), \quad o \in \{O_i, \emptyset\}. \tag{20}$$

For the original view,

$$\begin{aligned} P(Y_i \mid V_i^{(0)}, T_i) &= P(Y_i \mid X_i, O_i, T_i) \\ &= P(Y_i \mid X_i, T_i). \end{aligned} \tag{21}$$

For the ideal reference view,

$$\begin{aligned} P(Y_i \mid V_i^{(r)}, T_i) &= P(Y_i \mid X_i, \emptyset, T_i) \\ &= P(Y_i \mid X_i, T_i). \end{aligned} \tag{22}$$

Combining Eq. (21) and Eq. (22),

$$\begin{aligned} P(Y_i \mid V_i^{(0)}, T_i) &= P(Y_i \mid V_i^{(r)}, T_i) \\ &= P(Y_i \mid X_i, T_i). \end{aligned} \tag{23}$$

For any $z \in \{0, r\}$,

$$P(Y_i \mid Z = z, X_i, T_i) = P(Y_i \mid X_i, T_i). \tag{24}$$

Thus,

$$Y_i \perp Z \mid X_i, T_i, \qquad I(Y_i; Z \mid X_i, T_i) = 0. \tag{25}$$

A.2 Proof of Proposition 2

Let

$$b_i = |g_{\emptyset,i}|. \tag{26}$$

By Assumption 2,

$$g_i = \tau_i + \epsilon_i, \qquad |\epsilon_i| \le b_i, \qquad \epsilon_i \in [-b_i, b_i]. \tag{27}$$

The filtered change is

$$\Delta_i = \operatorname{sign}(g_i)\,[|g_i| - b_i]_+ . \tag{28}$$

For any scalar x and any $b \ge 0$,

$$\operatorname{dist}(x, [-b, b]) = [|x| - b]_+ . \tag{29}$$

Applying Eq. (29) with $x = g_i$ and $b = b_i$,

$$\begin{aligned} |\Delta_i| &= [|g_i| - b_i]_+ \\ &= \min_{e \in [-b_i, b_i]} |g_i - e|. \end{aligned} \tag{30}$$

Assume $|g_i| > b_i$. Then

$$\operatorname{sign}(\Delta_i) = \operatorname{sign}(g_i). \tag{31}$$

Since $\tau_i = g_i - \epsilon_i$,

$$\begin{aligned} \operatorname{sign}(g_i)\tau_i &= |g_i| - \operatorname{sign}(g_i)\epsilon_i \\ &\ge |g_i| - |\epsilon_i| \\ &\ge |g_i| - b_i \\ &> 0. \end{aligned} \tag{32}$$

Hence,

$$\operatorname{sign}(\tau_i) = \operatorname{sign}(\Delta_i). \tag{33}$$

By the reverse triangle inequality,

$$\begin{aligned} |\tau_i| &= |g_i - \epsilon_i| \\ &\ge |g_i| - |\epsilon_i| \\ &\ge |g_i| - b_i \\ &= |\Delta_i|. \end{aligned} \tag{34}$$

For the calibration path,

$$\begin{aligned} \tilde{s}_i^{(0)}(\alpha) &= s_i^{(0)} - \frac{\alpha}{2}\Delta_i, \\ \tilde{s}_i^{(r)}(\alpha) &= s_i^{(r)} + \frac{\alpha}{2}\Delta_i. \end{aligned} \tag{35}$$

The calibrated gap is

$$\begin{aligned} \tilde{s}_i^{(0)}(\alpha) - \tilde{s}_i^{(r)}(\alpha) &= s_i^{(0)} - s_i^{(r)} - \alpha\Delta_i \\ &= g_i - \alpha\Delta_i. \end{aligned} \tag{36}$$

We denote this calibrated gap by $\tilde{g}_i(\alpha) = \tilde{s}_i^{(0)}(\alpha) - \tilde{s}_i^{(r)}(\alpha)$. The score center is preserved:

$$\frac{\tilde{s}_i^{(0)}(\alpha) + \tilde{s}_i^{(r)}(\alpha)}{2} = \frac{s_i^{(0)} + s_i^{(r)}}{2}. \tag{37}$$

If $|g_i| \le b_i$, then $\Delta_i = 0$ and

$$[|\tilde{g}_i(\alpha)| - b_i]_+ = 0 = (1 - \alpha)|\Delta_i|. \tag{38}$$

If $|g_i| > b_i$, then

$$\Delta_i = \operatorname{sign}(g_i)(|g_i| - b_i), \tag{39}$$

and

$$\begin{aligned} \tilde{g}_i(\alpha) &= g_i - \alpha\Delta_i \\ &= \operatorname{sign}(g_i)\,[b_i + (1 - \alpha)(|g_i| - b_i)]\,. \end{aligned} \tag{40}$$

Therefore,

$$\begin{aligned} [|\tilde{g}_i(\alpha)| - b_i]_+ &= (1 - \alpha)(|g_i| - b_i) \\ &= (1 - \alpha)|\Delta_i|. \end{aligned} \tag{41}$$

A.3 Proof of Proposition 3

For fixed sample i and strength α,

$$Q_i^{(z)}(\alpha) = \mathrm{Bern}\left(\sigma(\tilde{s}_i^{(z)}(\alpha))\right), \quad z \in \{0, r\}. \tag{42}$$

Let R_i be a view-invariant predictor. Define

$$\mathcal{L}_i(\alpha, R_i) = \frac{1}{2}\,\mathrm{CE}\left(Q_i^{(0)}(\alpha), R_i\right) + \frac{1}{2}\,\mathrm{CE}\left(Q_i^{(r)}(\alpha), R_i\right). \tag{43}$$

Let

$$M_i(\alpha) = \frac{1}{2}Q_i^{(0)}(\alpha) + \frac{1}{2}Q_i^{(r)}(\alpha). \tag{44}$$

By linearity in the first argument of cross-entropy,

$$\mathcal{L}_i(\alpha, R_i) = \mathrm{CE}(M_i(\alpha), R_i). \tag{45}$$

Using the standard decomposition,

$$\mathrm{CE}(M_i(\alpha), R_i) = H(M_i(\alpha)) + \mathrm{KL}(M_i(\alpha) \| R_i). \tag{46}$$

Minimizing over R_i gives

$$\min_{R_i} \mathcal{L}_i(\alpha, R_i) = H(M_i(\alpha)). \tag{47}$$

By the definition of Jensen-Shannon divergence,

$$\begin{aligned} D_i(\alpha) &= \frac{1}{2}\,\mathrm{KL}\left(Q_i^{(0)}(\alpha) \| M_i(\alpha)\right) \\ &\quad + \frac{1}{2}\,\mathrm{KL}\left(Q_i^{(r)}(\alpha) \| M_i(\alpha)\right) \\ &= H(M_i(\alpha)) - \bar{H}_i(\alpha). \end{aligned} \tag{48}$$

Thus,

$$H(M_i(\alpha)) = D_i(\alpha) + \bar{H}_i(\alpha). \tag{49}$$

Combining Eq. (47) and Eq. (49),

$$\min_{R_i} \mathcal{L}_i(\alpha, R_i) = D_i(\alpha) + \bar{H}_i(\alpha). \tag{50}$$

Adding the quadratic penalty,

$$\begin{aligned} &\min_{R_i} \mathcal{L}_i(\alpha, R_i) + \lambda\alpha^2 \\ &\quad = D_i(\alpha) + \bar{H}_i(\alpha) + \lambda\alpha^2. \end{aligned} \tag{51}$$

Therefore,

$$\begin{aligned} &\arg\min_{\alpha \in \mathcal{A}} \left[D_i(\alpha) + \bar{H}_i(\alpha) + \lambda\alpha^2\right] \\ &\quad = \arg\min_{\alpha \in \mathcal{A}} \left[\min_{R_i} \mathcal{L}_i(\alpha, R_i) + \lambda\alpha^2\right]. \end{aligned} \tag{52}$$

B Prompt and Implementation Details

Classification instruction. We use the same direct classification instruction for all samples:

```
USER: <image>
Text: {external text}
Classify this meme as harmful or benign.
Answer with one word.
ASSISTANT:
```

Here, `<image>` denotes the visual input token, and `{external text}` is the external text associated with the meme.

Neutral-probe instruction. For the neutral probe, we keep the image input and the classification instruction unchanged, but replace the sample-specific external text with a fixed neutral text:

```
USER: <image>
Text: An image containing visual content.
Classify this meme as harmful or benign.
Answer with one word.
ASSISTANT:
```

Candidate label scoring. We use the scoring function defined in Section 3.2. The two class labels are `harmful` and `benign`. For each evaluated tokenizer, we map these label strings to their corresponding answer tokens in the assistant response position. Let w_H and w_B denote the answer tokens for `harmful` and `benign`, respectively. We set

$$W_H = \{w_H\}, \qquad W_B = \{w_B\}. \tag{53}$$

The harmfulness score is then computed by Eq. (2). We verify that both label strings are represented as single answer tokens for the evaluated tokenizers. The final prediction is harmful if $s_\theta(V, T) \geq 0$, and benign otherwise.

Reference-view construction. We use CRAFT to detect visible-text regions and LaMa for local inpainting. Detected boxes are expanded by a fixed margin, clipped to the image boundary, and merged into a binary text mask. LaMa is applied only inside this mask, while pixels outside the mask are kept unchanged. The same detection, mask expansion, and inpainting settings are used across all datasets and backbones.

Calibration. We search the fixed candidate set $\mathcal{A} = \{0, 0.1, \ldots, 1.0\}$ and set $\lambda = 0.1$ for all datasets and backbones. If multiple candidates obtain the same objective value, we choose the smallest α.

C Ablation Variant Definitions

All ablation variants use the same final decision rule: the predicted label is obtained by thresholding a calibrated original-view score.

Vanilla. The vanilla baseline uses the original-view score directly:

$$s_i^{\text{Vanilla}} = s_i^{(0)}. \tag{54}$$

Raw-gap calibration. This variant removes neutral-probe filtering and directly uses the raw original-reference gap:

$$\Delta_i^{\text{raw}} = g_i. \tag{55}$$

The final score is

$$s_i^{\text{Raw}} = s_i^{(0)} - \frac{\alpha_i^\star}{2}\Delta_i^{\text{raw}}. \tag{56}$$

Here, $\alpha_i^\star$ is selected with the same objective and candidate grid as TIDC, but with Δ_i^{raw} replacing Δ_i.

Fixed-strength calibration. This variant keeps the neutral-probe filtered change Δ_i, but uses a fixed calibration strength $\alpha = 1$ for all samples:

$$s_i^{\text{Fixed}} = s_i^{(0)} - \frac{1}{2}\Delta_i. \tag{57}$$

Full TIDC. The full method uses both neutral-probe filtering and sample-specific strength selection:

$$s_i^{\text{TIDC}} = s_i^{(0)} - \frac{\alpha_i^\star}{2}\Delta_i. \tag{58}$$

Ethical considerations

This work studies harmful meme detection using public benchmark datasets. Example memes may contain offensive or sensitive content and are used only for research and evaluation. The proposed method produces classification signals and should not be used as the sole basis for content moderation or enforcement decisions.