Controlled Gradient Optimization for Harmful Video Detection

Anonymous ACL submission

Abstract

Harmful video detection exhibits a fundamental asymmetry. The underlying *intent* is often subtle and highly context dependent, whereas *spurious cues* such as emotionally charged audio or visual effects are salient and easy to exploit. As a result, standard multimodal models tend to overfit dominant but unreliable modalities, allowing them to dominate optimization and degrade generalization. We propose Controlled Gradient Optimization (CGO), a task-aware training framework that explicitly regulates cross-modal gradient interactions to enforce semantic consistency. CGO mitigates reliance on isolated, non-generalizable features through three complementary mechanisms. First, it enforces directional alignment of gradients to promote coherent cross-modal learning. Second, it suppresses unreliable updates using perturbation-aware reweighting to reduce the influence of uncertain signals. Third, it harmonizes convergence dynamics across modalities to prevent optimization imbalance. Extensive experiments on three real-world benchmarks show that CGO consistently achieves state-of-the-art performance. Furthermore, it demonstrates strong robustness under modality missingness and distribution shifts, establishing a stable and reliable training paradigm for safety-critical harmful video detection.

1 Introduction

Detecting harmful videos, including misinformation, hate speech, and manipulative content, remains a fundamental challenge in multimodal learning (Jo et al., 2024; Edstedt et al., 2022; Ha et al., 2024). Unlike generic classification tasks where salient features often correlate with labels, *harmful video detection exhibits a fundamental asymmetry*. Harmful intent is typically subtle, implicit, and highly context dependent, whereas spurious cues such as emotionally charged background music, sensationalist text, or stylized visual effects are salient and easy to exploit.

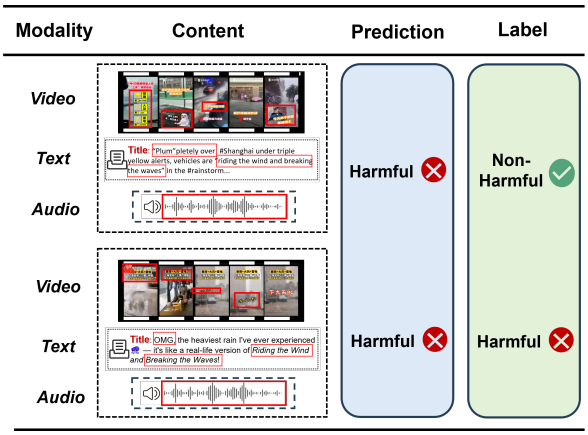


Figure 1: Typical failure cases in multimodal harmful video detection. Benign videos are misclassified as harmful when models over-rely on salient modality-specific cues, such as emotional text or stylized audio, despite weak harmful intent in the overall multimodal context.

This asymmetry creates a persistent optimization trap. Existing multimodal models prioritize signals that reduce training loss most rapidly, which causes them to overfit dominant but unreliable modality-specific cues (Gupta et al., 2023; Wang et al., 2025a; Zeng et al., 2025; Hussain et al., 2025; Zong et al., 2024, 2025). As illustrated in Figure 1, this leads to characteristic failure modes. Benign videos are misclassified as harmful due to loud but irrelevant features, while genuinely harmful content is missed because its subtle, distributed signals are overshadowed during training. Consequently, the primary bottleneck is not the lack of information, but the model's inability to resist shortcut learning driven by superficial salience (Geirhos et al., 2020; Zhou et al., 2021; Geirhos et al., 2018; Xiao et al., 2020; Wang et al., 2025b).

Addressing this challenge requires moving beyond simple feature fusion toward a task-aware op-

timization perspective. For a harmful prediction to be reliable, it should be supported by semantically consistent evidence across modalities, rather than being driven by a single high-magnitude gradient from a dominant modality.

To this end, we propose **Controlled Gradient Optimization (CGO)**, a training framework that explicitly regulates cross-modal gradient interactions to enforce semantic consistency. CGO fundamentally alters how multimodal information is aggregated by integrating three complementary mechanisms. First, **directional alignment** enforces geometric consistency among modality-specific gradients, ensuring that updates follow shared semantic directions rather than isolated noise. Second, **uncertainty suppression** uses perturbation-aware reweighting to down-weight modalities that exhibit high gradient variance. Third, **convergence harmonization** actively modulates learning dynamics to prevent dominant modalities from converging prematurely and overshadowing weaker but important signals.

Our main contributions are summarized as:

- **Optimization-Centric Problem Formulation:** We identify the asymmetry between subtle intent and salient noise as the root cause of training instability in harmful video detection, motivating a shift from architecture-centric to optimization-centric solutions.
- **The CGO Framework:** We introduce a unified training framework that explicitly regulates gradient directions and magnitudes, preventing spurious correlations from dominating the learning process.
- **Effectiveness and Robustness:** Extensive experiments on FakeSV, FakeTT, and HateMM show that CGO achieves superior detection accuracy while significantly improving robustness to modality loss and distribution shifts.

2 Related Work

Multimodal Fake News Video Detection. The rapid spread of fake news on short video platforms has driven growing interest in multimodal detection. Existing approaches integrate visual, audio, and textual information using techniques such as noise suppression (Qi et al., 2023), causal debiasing (Zeng et al., 2024), semantic and manipulation-aware modeling (Bu et al., 2024), and attention-based fusion (Zhang et al., 2025b). While these methods improve detection performance, they largely rely on feature fusion mechanisms that remain vulnerable to salient but spurious modality-specific cues, and often suffer from limited robustness under noisy or incomplete multimodal inputs.

Hateful Video Detection. Hateful video detection is challenging due to the implicit, context-dependent, and evolving nature of hateful expressions across modalities (Hee et al., 2024; Rehman et al., 2025; Koushik et al., 2025; Wang et al., 2025a). Prior work explores diverse strategies, including rationale-based reasoning (Lin et al., 2024), modality-aware composition (Cao et al., 2024), hierarchical co-learning (Wang et al., 2024a), and retrieval-augmented expert modeling (Lang et al., 2025). Despite encouraging progress, existing methods often struggle with subtle hateful intent expressed through weak or uneven multimodal signals, particularly when dominant but unreliable modalities overshadow semantically critical cues.

3 Optimization Motivation and Problem Formulation

3.1 Optimization Challenges from a Learning Dynamics Perspective

Harmful video detection poses a fundamental challenge for multimodal learning that goes beyond feature representation or fusion design. While harmful intent is typically subtle and implicitly distributed across modalities, misleading cues such as emotional text, stylized visuals, or rhythmic audio are often salient and locally predictive.

From a gradient-based optimization perspective, this asymmetry induces a systematic bias during training. Modalities that generate large-magnitude or low-variance gradients tend to dominate parameter updates, regardless of whether these signals are semantically reliable. As a result, multimodal models are prone to shortcut learning, where optimization is driven by isolated but salient modality-specific cues rather than coherent cross-modal evidence. This leads to characteristic failure modes, including false positives caused by emotionally charged but benign content, and false negatives where subtle harmful intent is overshadowed during training.

These observations suggest that the primary bottleneck of harmful video detection is not insufficient multimodal information, but the absence of mechanisms that regulate cross-modal learning dynamics. Effective optimization for this task there-

fore requires explicitly controlling how gradients from different modalities interact and contribute to parameter updates.

3.2 Expected Optimization State for Harmful Video Detection

Motivated by the above analysis, we formalize an *expected optimization state* θ^* that characterizes effective multimodal learning dynamics for harmful video detection. This state captures the conditions under which subtle harmful cues can be reliably aggregated across modalities, while preventing dominant but unreliable signals from driving optimization. Specifically, θ^* satisfies three properties:

Gradient Alignment. Modality-specific gradients should be directionally consistent:

$$\nabla f_i(\theta^*) \parallel \nabla f_j(\theta^*), \quad \forall i, j, \tag{1}$$

ensuring that parameter updates reflect semantically coherent cross-modal evidence rather than isolated modality reactions.

Gradient Stability. Gradient updates should exhibit low variance across training iterations:

$$\mathrm{Var}(\nabla f_i(\theta^*)) \to 0, \quad \forall i, \tag{2}$$

thereby suppressing noisy or unreliable modality signals that could destabilize optimization.

Modality Balance. The effective contributions of different modalities should remain balanced:

$$w_i(\theta^*) \approx \frac{1}{n}, \quad \forall i, \tag{3}$$

where w_i denotes the relative contribution of modality i and n is the number of modalities.

Together, these properties describe an optimization regime in which multimodal learning is guided by coordinated, stable, and balanced gradient dynamics, allowing subtle harmful intent to emerge from distributed cross-modal evidence.

3.3 Task Definition and Feature Extraction

We formulate multimodal harmful video detection as a supervised binary classification task. Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, each sample $x^{(i)} = (v^{(i)}, a^{(i)}, s^{(i)}, t^{(i)})$ consists of four synchronized modalities—video, audio, static image, and text—where $y^{(i)} \in \{0, 1\}$ indicates whether the content is harmful.

Following common practice in multimodal learning, each modality is processed using a two-stage pipeline: a frozen pretrained extractor to preserve general semantic priors, followed by a trainable encoder that adapts modality-specific representations to the harmful video detection task.

Specifically, textual inputs are obtained via OCR and subtitle parsing, and encoded using a BERT-based model to produce f_t. Static images are encoded using a frozen VGG19 backbone and a trainable image encoder, yielding f_i. Video inputs are processed by a frozen spatiotemporal encoder (e.g., ViT (Arnab et al., 2021) or C3D), followed by a trainable video encoder to obtain f_v. Audio signals are converted to spectrograms and encoded using a frozen VGGish or MFCC (Davis and Mermelstein, 1980) extractor together with a trainable audio encoder, producing f_a.

The resulting modality-specific embeddings $\{f_t, f_i, f_v, f_a\} \in \mathbb{R}^{B\times d}$ are concatenated and fed into a task-specific classifier to produce the prediction $\hat{y}^{(i)}$. The model is trained using the standard binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\big(y^{(i)}, \hat{y}^{(i)}\big). \tag{4}$$

4 Controlled Gradient Optimization

Guided by the expected optimization state defined in Section 3, we propose *Controlled Gradient Optimization (CGO)*, a task-aware training framework that explicitly regulates cross-modal learning dynamics. CGO operationalizes the three desired properties—gradient alignment, gradient stability, and modality balance—through three complementary components: GCC, PAD, and AOG, as illustrated in Figure 2.

4.1 GCC: Gradient-Constrained Consistency

To satisfy the gradient alignment property of the expected optimization state, CGO first enforces directional consistency among modality-specific gradients. In harmful video detection, emotionally expressive text or visually striking content often induces sharp but isolated gradients, which can dominate parameter updates despite weak semantic relevance.

Let $\{f_1, \ldots, f_n\}$ denote modality representations and $\{\nabla f_1, \ldots, \nabla f_n\}$ their corresponding gradients. We measure cross-modal gradient alignment using the mean pairwise cosine similarity:

$$\mathcal{A}_{\text{mean}} = \frac{2}{n(n-1)}\sum_{i<j} \cos(\nabla f_i, \nabla f_j), \tag{5}$$

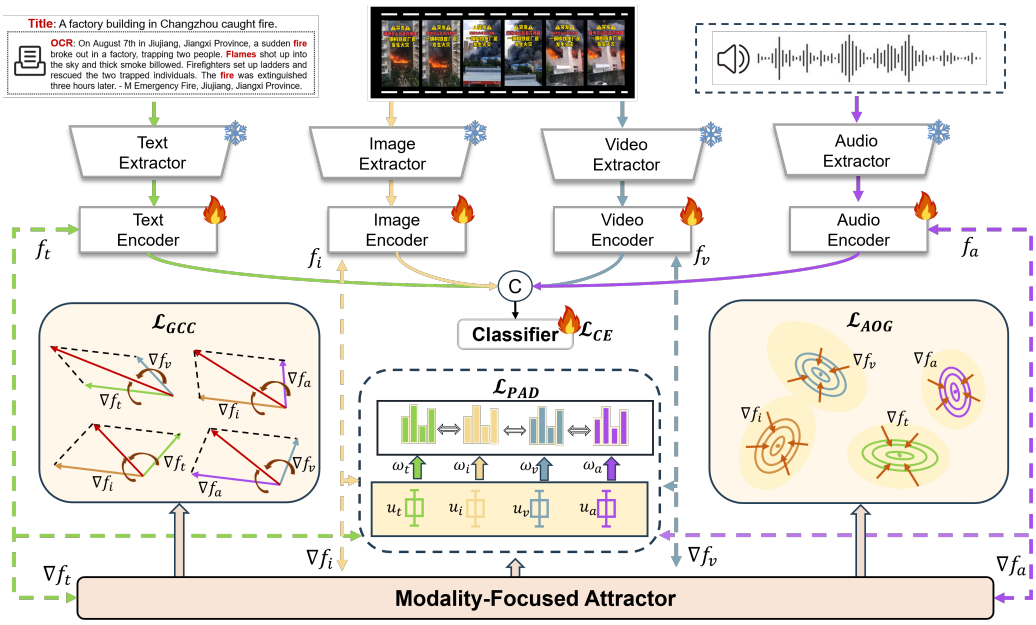


Figure 2: Overall Architecture of CGO: a task-aware gradient regulation framework for harmful video detection.

and define the gradient-constrained consistency loss as:

$$\mathcal{L}_{\text{GCC}} = 1 - \mathcal{A}_{\text{mean}}. \tag{6}$$

By explicitly encouraging gradients from different modalities to point toward a shared optimization direction, GCC prevents isolated modality reactions from driving learning and promotes semantically coherent cross-modal updates.

4.2 PAD: Perturbation-Aware Dynamic Reweighting

While gradient alignment ensures directional coherence, it does not account for the reliability of individual modality signals. In practice, modality-specific gradients may exhibit high variance due to noise, occlusion, or perturbations. This issue is particularly pronounced in harmful video detection, where OCR errors, background music, or visual effects can introduce spurious gradients.

To address this, we quantify modality uncertainty using gradient variance:

$$u_m = \text{Var}(\nabla f_m), \quad w_m = \frac{1}{1+u_m}, \tag{7}$$

and construct a perturbation-aware reweighted consistency loss:

$$\mathcal{L}_{\text{PAD}} = \sum_{i<j} \|w_i f_i - w_j f_j\|^2. \tag{8}$$

By down-weighting modalities with unstable gradients, PAD enforces the gradient stability property of the expected optimization state and suppresses unreliable updates that could distort harmful intent modeling.

4.3 AOG: Adaptive Optimization via Gradient Norm Modulation

Even with aligned and stable gradients, multimodal optimization can suffer from convergence imbalance when certain modalities dominate learning dynamics. Importantly, harmful video detection does not require equal modality importance, but it does require preventing dominant yet unreliable modalities from converging too rapidly and overshadowing weaker but semantically critical cues.

To harmonize convergence behavior, we adaptively modulate the learning rate for each modality based on relative gradient magnitudes:

$$\eta_m = \eta \left(1 + \sum_{n \neq m} \frac{\|\nabla f_n\|}{\|\nabla f_m\| + \epsilon}\right), \tag{9}$$

and regularize cross-modal convergence using:

$$\mathcal{L}_{\text{AOG}} = \sum_{m \neq n} \|\eta_m - \eta_n\|^2. \tag{10}$$

This mechanism enforces the modality balance property by harmonizing convergence rates across modalities, while preserving semantically meaningful asymmetry in their contributions.

4.4 Final Objective

The final training objective integrates task supervision with the proposed optimization constraints:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{GCC}} + \lambda_2 \mathcal{L}_{\text{PAD}} + \lambda_3 \mathcal{L}_{\text{AOG}}. \tag{11}$$

Together, these components guide the optimization process toward the expected state, ensuring that harmful predictions emerge from coordinated, stable, and balanced multimodal learning dynamics rather than isolated or emotionally salient signals.

5 Experiments

We conduct experiments to evaluate the effectiveness and robustness of **CGO** for multimodal harmful video detection, with a focus on mitigating modality imbalance and spurious cross-modal correlations that commonly degrade detection performance.

Our evaluation addresses the following research questions:

- **RQ1**: How does CGO compare with unimodal, multimodal, and large vision-language model baselines on harmful video detection benchmarks?
- **RQ2**: What are the contributions of individual components (GCC, PAD, and AOG) to performance and robustness?
- **RQ3**: How robust is CGO under missing-modality and distribution-shift scenarios?
- **RQ4**: Does CGO reduce modality dominance during training?

We evaluate CGO on three real-world datasets under both standard and degraded settings. In addition to detection performance, we analyze training behavior using *Modality Contribution Entropy* (MCE) as a diagnostic tool for understanding modality contribution during optimization.

5.1 Datasets

FakeSV (Qi et al., 2023) is a multimodal benchmark dataset designed for fake news detection on short video platforms. It comprises 3,654 short videos, evenly split between 1,827 fake and 1,827 real news samples. The videos were collected from platforms such as Douyin and Kuaishou, spanning the years 2017 to 2022.

FakeTT (Bu et al., 2024) is an English-language dataset tailored for fake news detection on short-video platforms. It includes 1,991 short videos, with 1,172 labeled as fake and 819 as real. The dataset features TikTok videos published between May 2019 and March 2024, each providing visual content, audio, and textual captions.

HateMM (Das et al., 2023) is a multimodal dataset for hate speech detection in short videos. It contains 1,083 videos from BitChute, with 431 labeled as hate and 652 as non-hate. Each video includes text transcripts, audio, and visual frames, along with annotated frame-level rationales and target communities.

5.2 Experimental settings

Training Details We train all models using the Adam optimizer with a batch size of 128 on NVIDIA 3090Ti GPUs. The hyperparameters λ_1, λ_2, and λ_3 are treated as trainable variables and jointly optimized with the model parameters. During training, these weights typically converge to values close to 0.1, allowing the model to adaptively balance task loss and auxiliary constraints. To ensure fair evaluation, we adopt five-fold cross-validation for all experiments.

Baselines for Fake News Video Detection. To verify the effectiveness of CGO, we compare it against competitive baselines from three categories: (1) *Unimodal detection methods* that use single modalities such as BERT (Radford et al., 2019), VGGish, VGG19 (Simonyan and Zisserman, 2014), and C3D (Tran et al., 2015); (2) *Multimodal fusion methods* that combine multiple modalities, including CAFE (Chen et al., 2022), TikTec (Shang et al., 2021), FANVN (Choi and Ko, 2021), HMCAN (Qian et al., 2021), SV-FEND (Qi et al., 2023), FakingRec (Bu et al., 2024), ExMRD (Hong et al., 2025), MMVD (Zeng et al., 2024) and PNRN (Kong et al., 2025); (3) *MLLM-based methods* which apply large vision-language models such as GPT-4o (Achiam et al., 2023), GPT-4, VideoLLaMA (Zhang et al., 2023) and Fact-R1 (Zhang et al., 2025a)

Baselines for Hateful Video Detection. For experiments on the HateMM dataset, we adopt a set of competitive multimodal hate detection models, including SharedCon (Ahn et al., 2024), Pro-Cap (Cao et al., 2023), HTMM (Das et al., 2023), RGCL (Mei et al., 2023), MHCL (Wang

Table 1: erformance comparison of CGO with unimodal models, traditional multimodal fusion methods, and large vision-language models on the FakeSV and FakeTT datasets.

	FakeSV				FakeTT			
Method	**ACC**	**M-F1**	**M-P**	**M-R**	**ACC**	**M-F1**	**M-P**	**M-R**
BERT	0.772	0.772	0.772	0.772	0.610	0.444	0.450	0.535
VGGish	0.716	0.715	0.718	0.716	0.642	0.499	0.465	0.574
VGG19	0.722	0.721	0.726	0.722	0.705	0.684	0.700	0.685
C3D	0.725	0.725	0.727	0.725	0.639	0.498	0.460	0.573
CAFE	0.662	0.657	0.673	0.663	0.655	0.630	0.645	0.634
TikTec	0.751	0.750	0.752	0.751	0.753	0.752	0.753	0.752
FANVN	0.750	0.750	0.751	0.750	0.762	0.746	0.762	0.742
HMCAN	0.728	0.725	0.739	0.729	0.681	0.621	0.709	0.639
SV-FEND	0.793	0.792	0.796	0.793	0.802	0.790	0.804	0.785
FakingRec	0.796	0.796	0.797	0.796	0.793	0.786	0.782	0.781
ExMRD	0.805	0.805	0.807	0.805	0.783	0.758	0.785	0.752
MMVD	0.826	0.826	0.827	0.826	0.804	0.793	0.804	0.790
PNRN	0.833	0.833	0.836	0.833	0.817	0.817	0.818	0.811
VideoLLaMA	0.588	0.584	0.591	0.596	0.663	0.487	0.519	0.547
GPT-4	0.746	0.743	0.740	0.760	0.622	0.622	0.689	0.681
GPT-4o	0.739	0.734	0.735	0.734	0.666	0.655	0.684	0.663
Fact-R1	0.756	0.747	0.777	0.720	0.744	0.727	0.778	0.683
CGO	**0.851**	**0.851**	**0.851**	**0.851**	**0.835**	**0.832**	**0.834**	**0.830**

et al., 2024a), Mod-HATE (Cao et al., 2024), ExplainHM (Lin et al., 2024), and MoRE (Lang et al., 2025), as well as MLLM-based methods like MiniCPM-V (Yao et al., 2024), LLaVA-OV (Li et al., 2024) and Qwen2-VL (Wang et al., 2024b).

Table 2: Experimental results of the competitive baseline models and the proposed CGO on the HateMM dataset.

Method	**ACC**	**M-F1**	**M-P**	**M-R**
SharedCon cite	0.696	0.686	0.687	0.685
Pro-Cap	0.645	0.633	0.634	0.632
HTMM	0.760	0.728	0.779	0.720
RGCL	0.756	0.736	0.730	0.752
MHCL	0.774	0.765	0.765	0.766
Mod-HATE	0.687	0.654	0.651	0.676
ExplainHM	0.732	0.689	0.682	0.701
MoRE	0.812	0.807	0.815	0.801
MiniCPM-V	0.724	0.723	0.778	0.764
LLaVA-OV	0.756	0.756	0.779	0.783
Qwen2-VL	0.737	0.737	0.781	0.773
CGO	0.825	0.812	0.817	0.809

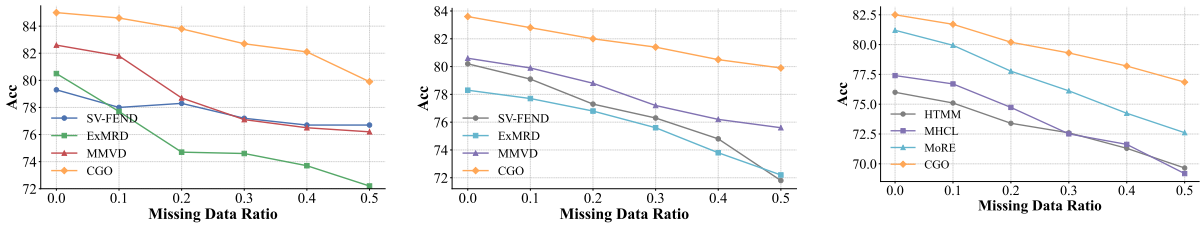
5.3 Performance on Harmful Video Detection (RQ1)

We evaluate the effectiveness of the proposed **CGO** framework on two representative harmful video detection tasks: fake news detection and hateful content classification. Experiments are conducted on three benchmarks—FakeSV, FakeTT, and HateMM—using Accuracy (ACC), Macro-F1 (M-F1), Macro-Precision (M-P), and Macro-Recall (M-R) as evaluation metrics, averaged over five-fold cross-validation.

Fake News Video Detection. On both FakeSV and FakeTT, CGO consistently outperforms all unimodal, multimodal, and large vision-language model baselines across all metrics. Compared with prior approaches that often overfit to stylistic patterns or emotionally salient cues (e.g., sensational titles or background music), CGO encourages the model to rely on semantically consistent evidence across modalities. As a result, CGO reduces false positives caused by isolated modality-specific signals and exhibits stronger generalization under diverse content styles.

Hateful Video Detection. CGO also achieves superior performance on the HateMM dataset. Hateful intent in short videos is frequently implicit and weakly expressed across modalities, making it difficult for fusion-based models to detect reliably. By suppressing the dominance of unreliable modalities and promoting coordinated learning across heterogeneous signals, CGO enables more effective integration of subtle multimodal cues, leading to improved detection of implicit hateful content.

Overall, the results demonstrate that CGO provides consistent performance gains across different harmful video detection scenarios. By mitigating modality dominance and encouraging semantically



(a) FakeSV (b) FakeTT (c) HateMM

Figure 3: Accuracy vs. Missing Data Ratio across three datasets.

Table 3: Ablation study on the contribution of key components in CGO across FakeSV, FakeTT, and HateMM datasets (ACC). We analyze the impact of removing specific modules.

Method	FakeSV	FakeTT	HateMM
w/o GCC	0.838	0.821	0.811
w/o PAD	0.842	0.822	0.805
w/o AOG	0.840	0.818	0.817
CGO	0.851	0.835	0.825

grounded multimodal learning, CGO achieves both higher accuracy and improved robustness, highlighting its effectiveness as a task-aware optimization framework for harmful video detection.

5.4 Ablation Study on Task-Aware Components (RQ2)

We conduct ablation experiments on FakeSV, FakeTT, and HateMM to examine the contribution of each component in **CGO**. Specifically, we remove gradient-consistent alignment (GCC), perturbation-aware reweighting (PAD), and adaptive optimization (AOG) from the final objective. Results are summarized in Table 3.

Effect of Gradient-Consistent Alignment (GCC). Removing GCC consistently degrades performance across all datasets, indicating that cross-modal gradient consistency is important for harmful video detection. Without GCC, the model becomes more sensitive to isolated modality-specific cues, leading to reduced accuracy and stability.

Effect of Perturbation-Aware Reweighting (PAD). Ablating PAD results in noticeable robustness drops, particularly under noisy or incomplete inputs. This suggests that suppressing unreliable modality signals is critical, as such cues can otherwise dominate learning and degrade detection performance.

Effect of Adaptive Optimization (AOG). Disabling AOG also leads to performance degradation, highlighting the importance of balancing convergence dynamics across modalities. Without adaptive optimization, dominant modalities tend to overshadow weaker yet semantically informative signals, which is detrimental for detecting subtle harmful intent.

Overall, the ablation results show that GCC, PAD, and AOG contribute in complementary ways. Their combination enables CGO to achieve more stable and robust learning, resulting in consistent performance gains across multimodal harmful video detection benchmarks.

5.5 Robustness to Missing Modalities (RQ3)

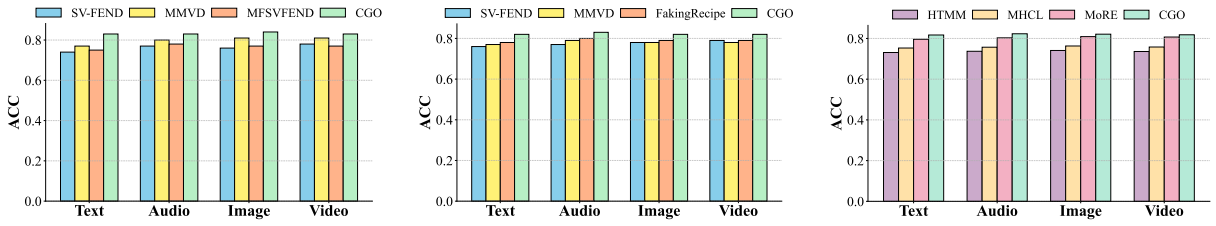
Random Modality Dropout. We evaluate robustness to incomplete inputs by randomly masking one or more modalities with dropout ratios ranging from 0.1 to 0.5. As shown in Figure 3, **CGO** consistently outperforms competing methods across all benchmarks, indicating reduced reliance on any single modality under missing-modality conditions.

Single-Modality Ablation. We further assess robustness by removing each modality individually. As illustrated in Figure 4, CGO exhibits minimal performance degradation across all ablation settings, suggesting that it avoids overfitting to dominant modalities and promotes coordinated multimodal representations.

Overall, these results demonstrate that CGO is robust to both random and targeted modality loss, supporting reliable harmful video detection under incomplete or degraded multimodal inputs.

5.6 Modality Contribution Entropy (RQ4)

To understand how CGO regulates modality contributions during training, we analyze *Modality Contribution Entropy (MCE)*. We stress that MCE is not



(a) FakeSV (b) FakeTT (c) HateMM

Figure 4: Robustness comparison of various models on three datasets, evaluated by accuracy under single-modality ablation.

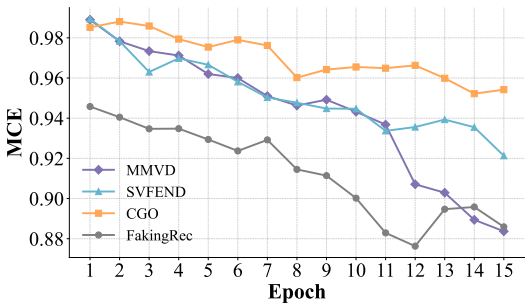


Figure 5: MCE curve on the FakeSV dataset. CGO sustains higher entropy throughout training, reflecting stable modality coordination and convergence toward the Modality-Focused Attractor.

a new evaluation metric for harmful video detection, but a diagnostic tool for interpreting gradient-level modality dominance during optimization.

Given K modalities with L2 gradient norms $\|\nabla f_1\|, \ldots, \|\nabla f_K\|$, we compute the normalized contribution distribution:

$$p_m = \frac{\|\nabla f_m\|}{\sum_{j=1}^{K} \|\nabla f_j\|}. \tag{12}$$

Here, p_m reflects the relative share of gradient-driven updates attributed to modality m. Compared to pairwise cosine similarity (which mainly captures directional agreement), $\{p_m\}_{m=1}^{K}$ summarizes the global dominance structure of optimization across all modalities.

MCE is defined as the entropy of this distribution, normalized by $\log K$:

$$\mathrm{MCE} = -\frac{1}{\log K} \sum_{m=1}^{K} p_m \log(p_m), \tag{13}$$

so that $\mathrm{MCE} \in [0, 1]$. Higher MCE indicates more balanced gradient contributions, while lower MCE implies optimization dominated by a few modalities, which can induce shortcut learning on spurious or unreliable cues.

We compute per-epoch MCE by averaging batch-level MCE values within each epoch. As shown in Figure 5, CGO maintains consistently higher MCE throughout training on FakeSV, indicating reduced modality dominance and more balanced optimization dynamics. This aligns with CGO's design: GCC promotes cross-modal gradient alignment to avoid isolated reactions, PAD down-weights unstable (high-variance) modalities, and AOG harmonizes convergence to prevent premature dominance. In contrast, competing methods exhibit rapid entropy decay, suggesting increasing reliance on a small subset of modalities, which is consistent with their inferior robustness under missing-modality and distribution-shift settings.

6 Conclusion

Multimodal harmful video detection is challenging due to the asymmetry between subtle harmful intent and salient but unreliable cues, which often induces optimization imbalance and shortcut learning in standard multimodal models. We propose Controlled Gradient Optimization (CGO), a task-aware framework that regulates cross-modal learning dynamics at the gradient level by aligning gradients, down-weighting unstable modalities, and harmonizing convergence to prevent premature modality dominance. Experiments on multiple benchmarks show consistent performance gains and improved robustness under missing modalities and distribution shifts, and Modality Contribution Entropy analysis further confirms reduced modality dominance and more balanced optimization dynamics during training. Overall, CGO demonstrates that optimization-centric modality regulation is a principled and effective paradigm for reliable harmful video detection.

Limitations

This study suffers limitations that may impact the performance of our proposed framework. While introducing retrieval-augmented contrastive learning strategies has achieved promising results in fake news detection, the performance of the retrieve samples may have the influence on the accuracy of fake news detection. Moreover, although the incomplete-modality-tolerant learning framework is effective in modeling cross-modal and cross-sample consistency for incomplete modalities imagination for multi-domain fake news video detection, extremely small prediction scores may result in an abundance of zero values, posing a risk of overfitting or gradient vanishing. We plan to address these limitations in future study.

Ethics Statement

This paper adheres to the ACM Code of Ethics and Professional Conduct. Firstly, the dataset utilized does not contain sensitive private information and poses no harm to society. Secondly, proper attribution is given to relevant papers and the sources of pre-trained models, along with detailed references to the toolkits used. Furthermore, our code will be released under the license of any artifacts used. Lastly, the proposed fake news video detection method is designed to contribute to the safety and stability of the internet environment and public opinion.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10444–10455.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 1351–1360, New York, NY, USA. Association for Computing Machinery.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, pages 4575–4584.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

Hyewon Choi and Youngjoong Ko. 2021. Using topic modeling and adversarial neural networks for fake news video detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2950–2954.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Johan Edstedt, Amanda Berg, Michael Felsberg, Johan Karlsson, Francisca Benavente, Anette Novak, and Gustav Grund Pihlgren. 2022. Vidharm: A clip based dataset for harmful content detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1543–1549. IEEE.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

Shrey Gupta, Pratyush Priyadarshi, and Manish Gupta. 2023. Hateful comment detection and hate target type prediction for video comments. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3923–3927.

Eungyeom Ha, Heemook Kim, and Dongbin Na. 2024. Hod: New harmful object detection benchmarks for robust surveillance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 183–192.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Lee. 2024. Recent advances in online hate speech moderation: Multimodality and the role of large models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 4684–4698.

Fiza Gulzar Hussain, Muhammad Wasim, Seemab Hameed, Abdur Rehman, Muhammad Nabeel Asim, and Andreas Dengel. 2025. Fake news detection landscape: Datasets, data modalities, ai approaches, their challenges, and future perspectives. *IEEE Access*.

Claire Wonjeong Jo, Magdalena Wojcieszak, and 1 others. 2024. Harmful youtube video detection: A taxonomy of online harm and mllms as alternative annotators. *arXiv preprint arXiv:2411.05854*.

Xiangzheng Kong, Zhi Zeng, Chenxi Zhu, Zihan Ma, and Minnan Luo. 2025. Harmony in chaos: A progressive noise-resilient network for robust fake news video detection. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Girish A Koushik, Diptesh Kanojia, and Helen Treharne. 2025. Towards a robust framework for multimodal hate detection: A study on video vs. image-based content. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2014–2023.

Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. 2025. Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection. In *Proceedings of the ACM on Web Conference 2025*, pages 2763–2774.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2023. Improving hateful meme detection through retrieval-guided contrastive learning. *arXiv preprint arXiv:2311.08110*.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mohammad Zia Ur Rehman, Anukriti Bhatnagar, Omkar Kabde, Shubhi Bansal, and Nagendra Kumar. 2025. Implihatevid: A benchmark dataset and two-stage contrastive learning framework for implicit hate speech detection in videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17209–17221.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Han Wang, Rui Yang Tan, and Roy Ka-Wei Lee. 2025a. Cross-modal transfer from memes to videos: Addressing data scarcity in hateful video detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5255–5263.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024a. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025b. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. 2024. Mitigating world biases: A multimodal multi-view debiasing framework for fake news video detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6492–6500.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025. Imol: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933.

Fanrui Zhang, Dian Li, Qiang Zhang, Jun Chen, Gang Liu, Junxiong Lin, Jiahong Yan, Jiawei Liu, and Zheng-Jun Zha. 2025a. Fact-r1: Towards explainable video misinformation detection with deep reasoning. *arXiv preprint arXiv:2505.16836*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Liyuan Zhang, Yang Yajing, Yan Yang, Yong Liu, Zhongyan Gui, Ruofan Li, and Hao Fei. 2025b. Mfsvfnd: Multimodal fusion network for detecting fake news on short video platforms. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 2123–2127.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. 2021. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pages 12857–12867. PMLR.

Linlin Zong, Wenmin Lin, Jiahui Zhou, Xinyue Liu, Xianchao Zhang, Bo Xu, and Shimin Wu. 2025. Text-guided fine-grained counterfactual inference for short video fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1237–1245.

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. 2024. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10817–10826.