Shedding the Facades, Connecting the Domains: Detecting Shifting Multimodal Hate Video with Test-Time Adaptation

Jiao Li[1, 2], Jian Lang[1], Xikai Tang[1], Wenzheng Shu[3], Ting Zhong[1*], Qiang Gao[4], Yong Wang[5], Leiting Chen[1, 2], Fan Zhou[1, 2]

[1]University of Electronic Science and Technology of China
[2]Intelligent Digital Media Technology Key Laboratory of Sichuan Province
[3]Kuaishou Technology
[4]Southwestern University of Finance and Economics
[5]Aiwen Technology Co., Ltd.
jiao_li@std.uestc.edu.cn, zhongting@uestc.edu.cn

Abstract

Hate Video Detection (HVD) is crucial for online ecosystems. Existing methods assume identical distributions between training (source) and inference (target) data. However, hateful content often evolves into irregular and ambiguous forms to evade censorship, resulting in substantial semantic drift and rendering previously trained models ineffective. Test-Time Adaptation (TTA) offers a solution by adapting models during inference to narrow the cross-domain gap, while conventional TTA methods target mild distribution shifts and struggle with the severe semantic drift in HVD. To tackle these challenges, we propose **SCANNER**, the first TTA framework tailored for HVD. Motivated by the insight that, despite the evolving nature of hateful manifestations, their underlying cores remain largely invariant (i.e., targeting is still based on characteristics like gender, race, etc), we leverage these stable cores as a bridge to connect the source and target domains. Specifically, SCANNER initially reveals the stable cores from the ambiguous layout in evolving hateful content via a principled centroid-guided alignment mechanism. To alleviate the impact of outlier-like samples that are weakly correlated with centroids during the alignment process, SCANNER enhances the prior by incorporating a sample-level adaptive centroid alignment strategy, promoting more stable adaptation. Furthermore, to mitigate semantic collapse from overly uniform outputs within clusters, SCANNER introduces an intra-cluster diversity regularization that encourages the cluster-wise semantic richness. Experiments show that SCANNER outperforms all baselines, with an average gain of 4.69% in Macro-F1 over the best.

1 Introduction

The explosive growth of social media has led to a surge in online videos for messaging, which unfortunately foster the creation of hateful multimodal (e.g., audio, text, and visual) content[1]. In particular, the dissemination of hateful content poses significant social challenges. Such hateful content targets specific groups (e.g., race, gender, nationality) to spread discrimination and violence, contributing to social discord and being widely amplified via online platforms (Das et al. 2023; Nan et al. 2025). Therefore, developing effective methods for Hate Video Detection (HVD) has become a pressing and urgent need.

(a) Explicit-style (b) Meme-style (c) Cartoon-style

Figure 1: Examples of evolving hateful manifestations. (a) indicates more explicit forms of hateful content, (b) and (c) represent more implicit and veiled manifestations.

Existing works in HVD adopt multimodal fusion models for detection (Zhang et al. 2023; Wang et al. 2024a). However, the manifestations of hateful content **evolve rapidly** (Lang et al. 2025), driven by creators' strategies to evade censorship, attract young audiences, and maximize user engagement. Specifically, overtly blunt hateful depictions are increasingly replaced by implicit or stylized manifestations, such as garbled or irregular language, meme-style edits, or cartoon imagery (cf. Figure 1). As a result, existing methods trained on historical data under the assumption of distributional consistency struggle to detect novel hateful content. Although MoRE (Lang et al. 2025) attempts to address this, it requires accessing both source (training) data and target (inference) labels, which is often unrealistic in real-world scenarios. Due to the extremely sensitive and harmful nature of hateful content, raw training data is often inaccessible and manual annotation is costly and slow, hindering timely detection. These challenges necessitate a Test-Time Adaptation (TTA) setting in HVD, where a model trained on the source domain is expected to adapt to an unseen target domain exhibiting distributional shifts, without

*Corresponding author.

[1]Disclaimer: This paper contains content that may be disturbing to some readers.

access to source data or target labels.

Conventional TTA methods employ self-supervised strategies like entropy minimization or pseudo-label self-training to implicitly bridge cross-domain gaps (Liang, Hu, and Feng 2020; Wang et al. 2021; Niu et al. 2022; Tsai et al. 2024). They mainly address mild and regular distributional shifts on originally clean data, such as variations in style or composition. However, the evolution of hateful content is essentially *an irregular, intent-driven shift in manifestations*, causing substantial yet uncertain semantic drift. As presented in Figure 1, while depicting the same hateful topic (i.e., racial discrimination), the left panel shows an explicit manifestation of hateful content, overtly disseminating harmful messages through direct visual symbols and aggressive rhetoric like "WHITE POWER ve NIGGER". In contrast, the center and right panels present ambiguous hateful manifestations, conveyed via humorous or cartoon-style visuals and abstract language like "GUARO CHILD Double all up ". As a result, how to perform effective TTA for HVD under severe semantic drift is still an open issue.

To address this, we observe that although the evolving manifestations of hateful content are ceaseless, the hateful cores, such as the targeted demographic categories (e.g., gender, race), remain largely constant. As shown in Figure 1, the targeted groups in all examples predominantly belong to specific racial categories, even though the manifestations of hateful expression vary substantially. *These invariant hateful cores*, learned and internalized by the source model during training, *can serve as a bridge to connect the source and target domains*. Drawing upon these observations, a straight thought for tackling the evolving manifestations is to isolate the invariant hateful cores and align target samples toward these cores, reducing sensitivity to surface-level variations while preserving focus on the underlying semantic intent.

Motivated by these insights and the limitations of existing approaches, we design a novel TTA paradigm tailored for HVD. It aims to guide target samples toward the source domain by eliminating superficial manifestations surrounding the hateful cores and revealing underlying harmful intent. Initially, to reduce sensitivity to new manifestations, we propose a Centroid-guided AligNment framework, **CAN**, which strengthens the source-target connection by first clustering target domain samples into hateful centroids and then encouraging sample alignment toward these invariant cores. Besides, to prevent the forced alignment of outlier-like samples that are weakly associated with centroids from disrupting gradient optimization (cf. Figure 3), we further extend CAN to an advanced Sample-adaptive Centroid AligNment framework, dubbed **SCAN**. SCAN dynamically assigns target instances with sample-aware alignment weights based on their affinity to the corresponding centroids, effectively stabilizing the adaptation. Furthermore, SCAN induces semantic collapse by homogenizing target features around centroids, resulting in overly uniform outputs within clusters (cf. Figure 4). Therefore, we finally upgrade SCAN into **SCANNER**, a Sample-adaptive Centroid AligNment framework with iNtra-clustER collapse avoidance. SCANNER incorporates an intra-cluster diversity regularization that promotes cluster-wise semantic diversity of target outputs by encouraging distinct class predictions within clusters. Our contributions are as follows:

- SCANNER is the pioneering TTA framework specifically designed for HVD, aiming to detect unseen and evolving hateful content in videos under drastic and unpredictable semantic drift in manifestations.
- SCANNER introduces a series of progressively refined alignment frameworks that guide manifestation-evolving hateful samples toward invariant harmful cores, while attenuating the adverse impact of cluster-distant outliers and enhancing intra-cluster semantic diversity.
- Extensive experiments on three datasets validate the superior adaptation performance of SCANNER, which significantly outperforms all baselines in achieving test-time HVD. The source codes are available at https://github.com/Jolieresearch/SCANNER.

2 Related Work

2.1 Hate Video Detection

Multimodal malicious content detection is the task of identifying malicious/hateful content expressed through multiple modalities. Early efforts focused on memes, leveraging pre-trained vision-language models to analyze their textual and visual components and capture nuanced cross-modal interactions (Cao et al. 2023b; Mei et al. 2025). Recently, researchers have turned to more challenging Hate Video Detection (HVD), which aims to identify hateful content embedded in visual, textual, and audio modalities within videos. They leveraged pre-trained encoders to extract modality features, subsequently utilizing fusion architectures to model their interactions for detection (Das et al. 2023; Wang et al. 2024a; Koushik, Kanojia, and Treharne 2025; Lang et al. 2025; Hong et al. 2025). However, such methods rely on the assumption of distributional consistency, which is violated in real-world settings where hateful content evolves rapidly, leading to significant performance degradation in detection. Although pivotal work like MoRE (Lang et al. 2025) begins to address distribution shifts caused by content evolution, it still relies on source data and target labels, resulting in untimely detection in the real world. In contrast, we present SCANNER for HVD under the source-free and target label-free setting, which solely leverages unlabeled target samples to align with the underlying semantics of hateful content by peeling off their superficial manifestations, enabling timely and robust adaptation to evolving hateful content.

2.2 Test-Time Adaptation

Test-Time Adaptation (TTA) aims to bridge the gap between source and target domains through unsupervised fine-tuning the source model using target domain data exhibiting a distribution shift. Tent updates the channel-wise affine parameters of batch normalization layers by entropy minimization (EM) (Wang et al. 2021). Subsequent research has proposed various EM variants to address challenges such as label shift and continual distribution changes (Wang et al. 2022; Zhou et al. 2023; Gan et al. 2023; Niu et al. 2023; Liu et al. 2024; Wang et al. 2024b; Zhou et al. 2025). Recently, some works

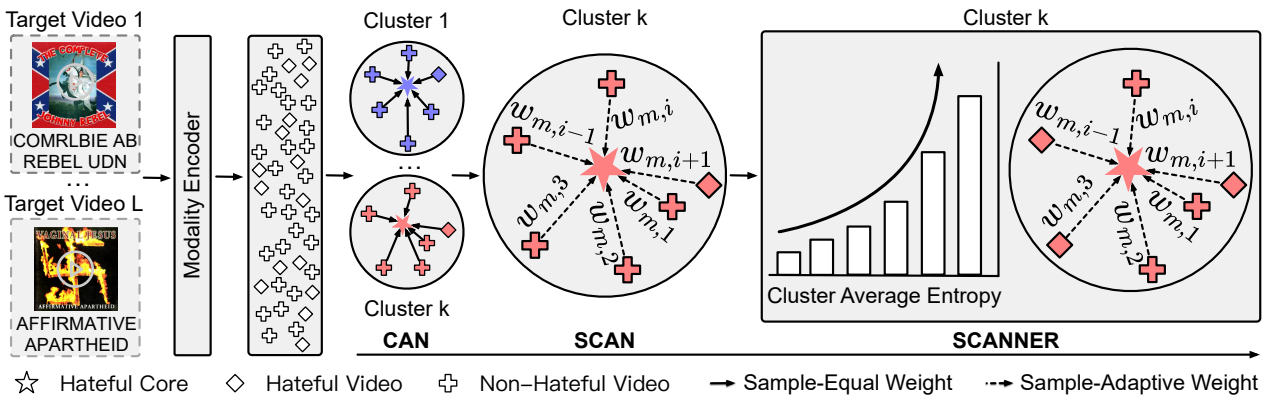


Figure 2: An overview of the SCANNER framework, illustrating its progressive evolution from the base framework (CAN) to the final architecture (SCANNER). Video samples are grouped into different clusters (with two example clusters shown; different colors denote different clusters). Solid lines indicate equal weighting for all samples; dashed lines denote similarity-based adaptive weighting, with longer lines implying larger distances from the cluster core and correspondingly smaller weights.

have delved into the multimodal test-time adaptation (Cao et al. 2023a; Guo et al. 2025; Dong, Chatzi, and Fink 2025; Guo and Jin 2025). For example, READ mitigates reliability bias by designing an optimization objective based on EM to narrow the gap between source and target domains (Yang et al. 2024). However, most TTA methods are confined to handling standard distribution shifts (e.g., corruption, stylization) for implicitly aligning source and target domains, which are inadequate for addressing irregular and unpredictable manifestation evolution in hateful content. Consequently, to effectively mitigate the semantic drift in HVD, we propose a novel TTA approach, SCANNER. Our method explicitly leverages the invariant hateful cores (e.g., demographic categories like gender or race) as a bridge to align the source and target domains, effectively narrowing the cross-domain gap.

3 Methodology

3.1 Preliminary

HVD Definition. Let $\mathcal{D} = \{D_1, \cdots, D_N\}$ denote a HVD dataset, where N is the number of videos. Each video D_i comprises visual v_i , textual t_i, and audio a_i modalities: $D_i = (v_i, t_i, a_i)$. HVD aims to determine whether a video D_i contains hateful content (i.e., hateful or non-hateful) (Das et al. 2023; Lang et al. 2025).

Test-Time Adaptation Setting. A source model f_{θ_s} is trained on a source domain dataset $\mathcal{D}_s = \{D_i^s\}_{i=1}^N$ to learn a function $f_{\theta_s} : (D_i^s) \rightarrow \hat{y}_i$, where $D_i^s = (v_i^s, t_i^s, a_i^s, y_i^s)$. f_{θ_s} consists of multimodal feature encoders $f_m(\cdot)$, where $m \in \{v, t, a\}$, a transformer-based fusion module T_s (Tsai et al. 2024; Guo et al. 2025) and a classifier C_s. To mitigate the semantic drift in HVD, we adapt the source model f_{θ_s} on target domain datasets with significant distribution shift caused by the evolving hateful content within videos $\mathcal{D}_t = \{D_i^t\}_{i=1}^L = (v_i^t, t_i^t, a_i^t)_{i=1}^L$ for better detection, under the source-free and target label-free setting.

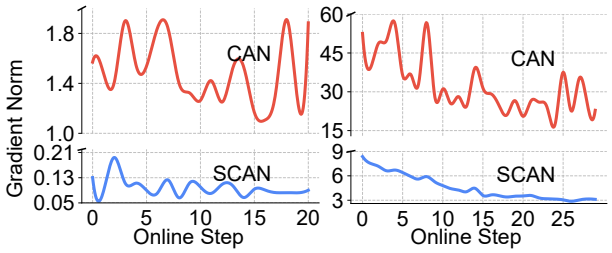
Overview Framework. To better cope with semantic drift in HVD, we introduce SCANNER. Figure 2 presents an overview of SCANNER framework and illustrates its progressive development: starting from the foundational framework CAN (Section 3.2), which encourages target samples to align with invariant hateful cores; extending to SCAN (Section 3.3), which incorporates a sample-level adaptive centroid alignment mechanism into CAN to facilitate sample-level adaptive weighting for more stable alignment; and culminating in SCANNER (Section 3.4), which further introduces an intra-cluster diversity regularization to prevent semantic collapse during adaptation.

3.2 CAN: Centroid-Guided Alignment

To adapt the source model for detecting evolving hateful videos at test time, we observe that the underlying hateful cores, such as targeting specific races, genders, or promoting violence, remain largely consistent across the source domain and target domain. These consistent hateful cores serve as a bridge to narrow the domain gaps, since the source model has already learned to detect such persistent hateful intents during pre-training, but only struggles to generalize to their novel surface manifestations. In light of these observations, we propose to alleviate the side effect of these novel and ambiguous manifestations during the adaptation by exposing their underlying hateful cores beneath these evolving forms to the source model. As a result, we initiate a **C**entroid-guided **A**lig**N**ment framework, termed **CAN**. The intuitive objective of CAN is to cluster unlabeled target videos and align them with stable centroids — representing invariant hateful cores — to encourage the pre-trained source model to focus on underlying harmful semantics it can recognize with ease, rather than superficial variations.

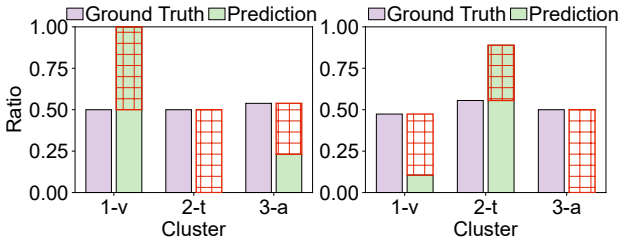
Concretely, for each modality m, where $m \in \{v, t, a\}$, we apply K-Means clustering (Hamerly and Elkan 2003; Liang, Hu, and Feng 2020) to partition the unlabeled target set $\mathcal{D}_t = \{D_i^t\}_{i=1}^L$ into k clusters at the modality level, forming a set of modality-specific centroids $C_m = \{c_{m,j}\}_{j=1}^k$:

$$c_{m,j} = \frac{1}{|S_{m,j}|} \sum_{D_i^t \in S_{m,j}} f_m(m_i^t), \tag{1}$$



(a) From MHY to HMM. (b) From MHY to MHB.

Figure 3: Evolution of gradient norm on CAN and SCAN frameworks during online adaptation. Source domain is MHY dataset, target domain are HMM and MHB datasets.



(a) From MHY to HMM. (b) From MHY to MHB.

Figure 4: Comparison of ground truth and predicted hateful video ratios (hateful vs. total videos) for one cluster per modality (1:visual, 2:textual, 3:audio) in SCAN. Grid regions indicate over- or under-estimation in predicted ratios relative to ground truth.

where $S_{m,j}$ denotes the set of the target video D_i^t from $\mathcal{D}_t$ assigned to the j-th cluster for modality m. To ensure the stability of the centroids C_m, we employ a momentum update strategy. For the τ-th online adaptation batch, the centroid $c_{m,j}^{(\tau)}$ is updated as follows:

$$c_{m,j}^{(\tau)} = \gamma \cdot c_{m,j}^{(\tau-1)} + (1-\gamma) \cdot c_{m,j}, \tag{2}$$

here $c_{m,j}^{(\tau-1)}$ is the centroid from the previous batch and $\gamma \in [0,1)$ is the momentum coefficient. Consequently, for each target video D_i^t and modality m, we compute its maximum similarity score $s_{m,i}$ to the k centroids, defined as follows:

$$s_{m,i} = \max_{j \in \{1,\dots,k\}} \left(\frac{f_m(m_i^t)^\top c_{m,j}^{(\tau)}}{\|f_m(m_i^t)\| \cdot \|c_{m,j}^{(\tau)}\|} \right), \tag{3}$$

where $c_{m,j}^{(\tau)}$ is the j-th centroid for batch τ. $\mathcal{L}_{\text{CAN}}$ is a basic clustering loss, which enforces intra-cluster compactness by encouraging each video to be close to their assigned centroids. $\mathcal{L}_{\text{CAN}}$ can be expressed concisely as the sum of losses over all modalities:

$$\mathcal{L}_{\text{CAN}} = \sum_{m \in \{v,t,a\}} \left(1 - \frac{1}{B} \sum\nolimits_{i=1}^{B} s_{m,i} \right), \tag{4}$$

here B is the number of videos in a mini-batch. Consequently, minimizing $\mathcal{L}_{\text{CAN}}$ effectively encourages the alignment of target video representations toward their corresponding centroids that capture the invariant hateful cores, thereby leading to tighter and more robust clustering despite the evolving manifestations of hateful videos.

3.3 SCAN: Sample-Adaptive Alignment

In the prior CAN framework, aligning each sample equally to respective centroid would forcibly pull some outlier-like samples that are weakly associated with centroids, leading to unstable gradient optimization that hurts the adaptation. To verify this hypothesis, we empirically investigate gradient norm behaviour by tracking the l_2-norm of all trainable parameter gradients during the optimization of CAN, as shown in Figure 3. The figure clearly shows that the gradient norm of CAN (red line) fluctuates drastically and spikes to abnormally high values, indicating a model collapse.

To get out of this pitfall, we upgrade CAN to a **S**ample-adaptive **C**entroid **A**lig**N**ment framework (**SCAN**), which evaluates the similarity between each target sample and its assigned centroid, and subsequently down-weights potential outliers. Specifically, we advance the loss function $\mathcal{L}_{\text{CAN}}$ into a sample-adaptive version $\mathcal{L}_{\text{SCAN}}$ by introducing a sample-adaptive weight $w_{m,i}$ for each target video D_i^t:

$$w_{m,i} = \text{softmax}(\beta \cdot s_{m,i}) = \frac{\exp(\beta \cdot s_{m,i})}{\sum_{j=1}^{B} \exp(\beta \cdot s_{m,j})}, \tag{5}$$

where β is a temperature parameter controlling the smoothness, and $s_{m,i}$ is the maximum similarity score. Finally, we formulate the sample-adaptive alignment loss:

$$\mathcal{L}_{\text{SCAN}} = \sum_{m \in \{v,t,a\}} \left(1 - \sum_{i=1}^{B} w_{m,i} \cdot s_{m,i} \right). \tag{6}$$

$\mathcal{L}_{\text{SCAN}}$ ensures that the source model prioritizes the alignment of centroid-close samples and suppresses the adverse effects of outliers Empirical results in Figure 3 show that, comparing to CAN, SCAN (blue line) effectively stabilizes gradient norms of alignment within a reasonable range.

3.4 SCANNER: Collapse-Avoidance Alignment

While SCAN mitigates the influence of outliers, it may incur homogenizing outputs of target samples (i.e., all samples are predicted as hateful or non-hateful) when aligning modality-specific representations with their respective centroids. As illustrated in Figure 4, we report the distributions of both ground truth and predictions produced by SCAN framework on three selected clusters across two target domains (HMM and MHB), each cluster corresponding to a distinct modality (visual, textual, and audio). Although the ground truth distributions within each modality-specific cluster are balanced (with ratios close to 0.5), the corresponding prediction distributions are significantly imbalanced (with ratios close to 0 or 1), revealing that SCAN tends to assign the majority of target samples within a modality-specific cluster to a single, incorrect class, despite differences in their labels. This phenomenon reflects a shortcut solution learned by the model

during alignment, making the model prone to modality-level semantic collapse (Hoang, Vo, and Do 2024), where it generates overly homogeneous predictions around the centroids.

Therefore, to address this issue of the aforementioned semantic collapse, we incorporate an intra-cluster diversity loss into the SCAN framework, thereby developing the **S**ample-adaptive **C**entroid **A**lig**N**ment framework with i**N**tra-clust**ER** collapse avoidance (**SCANNER**). SCANNER encourages high average prediction entropy at the cluster level, thereby promoting output diversity across samples within the same cluster. We compute the average prediction probability $\bar{p}^t_{m,n}$ of the n-th cluster in modality m:

$$\bar{p}^t_{m,n} = \frac{1}{|S_{m,n}|} \sum_{i \in S_{m,n}} \text{softmax}(o^t_{m,i}), \tag{7}$$

where $o^t_{m,i}$ is the logit output for target sample D^t_i, and $S_{m,n}$ is the set of videos assigned to the n-th cluster for modality m, $|S_{m,n}|$ represents the number of videos in the n-th cluster. We then formulate the diversity loss $\mathcal{L}_{\text{DIV}}$ as follows:

$$\mathcal{L}_{\text{DIV}} = \sum_{m \in \{v,t,a\}} \left(\frac{1}{k} \sum_{n=1}^{k} \left(\sum_{c=1}^{C_{cls}} \bar{p}^t_{(m,n),c} \cdot \log(\bar{p}^t_{(m,n),c}) \right) \right), \tag{8}$$

where C_{cls} is the number of classes and $\bar{p}^t_{(m,n),c}$ is the average prediction probability for class c over the n-th cluster for modality m. $\mathcal{L}_{\text{DIV}}$ encourages the average prediction distribution within each cluster to be as diverse as possible. And results from Figure 6 demonstrate that, in contrast to SCAN, SCANNER effectively mitigates the semantic collapse.

3.5 Adaptation Objective

During the test time, the source model parameters are updated using the following objective:

$$\mathcal{L} = \epsilon \cdot \mathcal{L}_{\text{EM}} + \lambda \cdot \mathcal{L}_{\text{SCAN}} + \alpha \cdot \mathcal{L}_{\text{DIV}}, \tag{9}$$

here ϵ, λ, and α are hyperparameters. Following prior TTA methods (Niu et al. 2022), we add an entropy minimization loss $\mathcal{L}_{\text{EM}}$ to encourage the source model to make more confident and clearer predictions on target samples during adaptation. These objectives work in tandem to enhance the source model's robustness to tackle the drastic semantic drift in evolving hateful videos, thereby improving generalization.

4 Experiments

4.1 Experimental Setup

Datasets. As shown in Table 1, we conduct experiments on three benchmarks in HVD: HateMM (HMM) (Das et al. 2023), MultiHateClip-Youtube (MHY), and MultiHateClip-Bilibili (MHB) (Wang et al. 2024a).

Setting of TTA. To evaluate the adaptability of our proposed SCANNER under real-world distribution shifts in HVD, we conduct cross-domain detection experiments. Each dataset originates from a distinct platform (BitChute, YouTube, Bilibili), differs in language (English or Chinese), and exhibits unique class distributions, leading to substantial semantic drift, as summarized in Table 1. These inherent discrepancies allow each dataset to be treated as a distinct domain, simulating semantic drift caused by the evolution of hateful videos. As a result, we define six evaluation groups, where each source-to-target domain adaptation setting is denoted as **Dataset A** $\rightarrow$ **Dataset B**.

Dataset	Total	Hateful(%)	Platform	Language
HMM	1,083	39.8	BitChute	English
MHY	1,000	33.8	YouTube	English
MHB	1,000	32.2	Bilibili	Chinese

Table 1: Characteristics of three hateful video datasets.

Baselines. Baselines are categorized into two classes: (1) Traditional HVD methods, which are pre-trained on the source domain dataset without adaptation, including HTMM (Das et al. 2023), MHCL (Wang et al. 2024a), ProCap (Cao et al. 2023b), and MoRE (Lang et al. 2025). (2) TTA methods, which adapt the source model to the unlabeled target data without accessing the source data, including Source, Test-time normalization (Norm) (Schneider et al. 2020), Self-Training (ST) (Liang, Hu, and Feng 2020), TENT (Wang et al. 2021), SAR (Niu et al. 2023), READ (Yang et al. 2024), and SuMi (Guo and Jin 2025).

Evaluation Metrics. Following prior works (Lang et al. 2025), we adopt two metrics to evaluate the model's performance: Classification Accuracy (ACC), Macro-F1 (M-F1).

Implementation Details. During the feature extraction phase, when the source and target domains are in the same language, we utilize the pre-trained CLIP (Radford et al. 2021) for feature extraction. In other cases, we employ multilingual Sentence-BERT (Reimers and Gurevych 2019) to ensure effective cross-lingual representation extraction. Following previous works (Ma 2024; Guo et al. 2025), we use transformer (Vaswani et al. 2017) as the source models. During the online test-time adaptation phase, we use AdamW optimizer, with an initial learning rate of 1.0×10^{-3}, weight decay of 5×10^{-4} and batch size of 128. The overall process is conducted within a single epoch. For learnable parameters, we update affine parameters in normalization layers and the linear layer in modal-specific encoders by following the prior works (Niu et al. 2023; Yang et al. 2024).

4.2 Main Performance

We report the performance in Table 2. From the results, we have the following observations:

O1: Superiority of TTA methods over traditional HVD approaches. Traditional HVD methods typically assume that the training (source) and inference (target) data share the same distributions, i.e., the assumption of distribution consistency. Therefore, these models fail drastically when faced with large distributional drift caused by variations in hateful manifestations. In contrast, TTA methods bridge the domain gap by employing entropy minimization (EM) to address mild distribution shifts. Therefore, they outperform traditional HVD methods under such conditions.

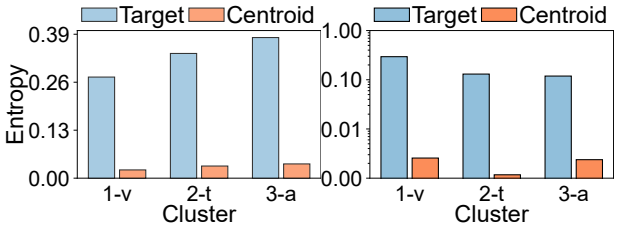
O2: SCANNER outperforms baseline methods in overall performance. Existing TTA methods are designed to cope with mild and structured distribution shifts. However, in the context of HVD, where target-domain hateful videos appear

Method	MHY→HMM		MHY→MHB		HMM→MHB		HMM→MHY		MHB→HMM		MHB→MHY		Avg.	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1
HTMM	59.46	57.33	56.00	50.55	58.30	50.50	64.80	47.50	45.71	45.56	53.00	52.11	56.21	50.59
MHCL	57.43	52.72	60.50	51.79	63.90	52.18	61.50	48.18	52.44	52.25	61.90	51.39	59.61	51.42
Pro-Cap	50.05	47.04	63.90	53.94	66.20	45.99	59.60	51.24	54.57	44.83	60.20	48.41	59.09	48.58
MoRE	55.03	54.71	65.70	44.39	37.10	32.87	50.20	50.13	51.25	36.87	61.30	43.34	53.43	43.72
Source	62.94	56.81	64.00	57.38	67.00	45.02	57.34	57.28	53.94	53.64	63.60	59.00	61.47	54.86
Norm	63.62	57.14	66.00	57.65	58.50	53.03	64.90	59.36	60.75	53.24	65.00	54.85	63.13	55.88
ST	60.66	60.39	66.00	54.17	**67.60**	40.33	65.00	40.97	52.72	52.72	66.10	39.83	63.01	48.07
TENT	63.06	56.19	65.80	59.81	58.00	53.01	64.30	60.47	61.03	54.02	66.30	58.07	63.08	56.93
SAR	63.67	59.26	65.65	59.73	58.10	53.15	65.20	61.28	61.23	55.70	66.10	57.91	63.32	57.84
READ	63.34	57.21	65.21	57.30	58.40	53.11	64.20	60.44	60.84	53.64	65.00	55.07	62.83	56.13
SuMi	63.34	56.85	65.60	56.82	57.00	52.23	63.70	59.75	60.94	53.63	66.20	56.05	62.80	55.89
SCANNER	**68.14**	**64.63**	**68.30**	**60.57**	60.20	**56.49**	**66.70**	**62.90**	**62.69**	**58.59**	**66.89**	**60.10**	**65.49**	**60.55**

Table 2: Performance comparisons between baselines and our SCANNER across six scenarios derived from the MHY, MHB, and HMM datasets. Bold and underlined values indicate the **best** and second-best performances, respectively.

Framework			MHY→HMM		MHY→MHB	
CAN	SACN	SCANNER	Acc	M-F1	Acc	M-F1
✗	✗	✗	62.94	56.81	64.00	57.38
✓	✗	✗	66.45	62.17	65.40	58.57
✗	✓	✗	67.74	64.02	65.72	58.95
✗	✗	✓	**68.14**	**64.63**	**68.30**	**60.57**

Table 3: Ablation studies comparing different variants of the framework. The best results are in black **bold**.



(a) From MHY to HMM. (b) From MHY to MHB.

Figure 5: Entropy comparison between modality-specific cluster (1:visual, 2:textual, 3:audio) centroids and the average entropy of target videos within corresponding cluster across two target domains.

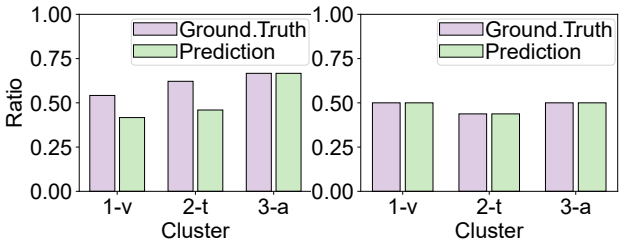
in highly diverse and irregular surface manifestations, resulting in a huge semantic drift that makes TTA methods fall short. Our proposed SCANNER addresses these challenges by explicitly aligning target videos to stable centroids. This reduces the source model's sensitivity to superficial variations in manifestations and effectively bridges the distribution gap in source and target domains. In addition, traditional methods tend to exhibit bias under severe class imbalance by over-fitting to the majority class, whereas our SCANNER explicitly encourages correct prediction of minority-class samples through centroid-guided alignment, demonstrating strong robustness against over-fitting.

4.3 Ablation Study

To further understand the effectiveness of each progressively upgraded variant of our frameworks, we conduct ablation studies. The performance of each framework is summarized in Table 3. The base framework **CAN** demonstrates improved detection performance over direct inference using the source model. This is attributed to its centroid-guided alignment, which aligns targets to corresponding invariant hateful centroids, thereby reducing sensitivity to diverse hateful manifestations. Building on this, **SCAN** incorporates a sample-adaptive alignment strategy that mitigates the adverse impact of low-similarity samples during alignment, thereby improving detection performance and stabilizing optimization. To address intra-cluster homogenization — where representations collapse toward the centroid — an intra-cluster diversity loss is introduced into the SCAN framework, resulting in the final framework **SCANNER**. This design preserves representational diversity within clusters and yields superior detection performance.

4.4 Centroids Stability Evaluation

To assess the rationality of the centroid-guided alignment in SCANNER, we select one cluster from each modality. For each modality-specific cluster, we compute and compare the prediction entropy of its centroid with the average prediction entropy of the associated target videos. Figure 5 reveals that the source model yields high entropy for target samples, intuitively reflecting a notable domain gap. This high-entropy phenomenon indicates that the source model struggles with confident predictions for target videos due to novel and diverse manifestations. However, the corresponding modality-specific centroids formed by clustering these target samples exhibit significantly lower prediction entropy. These low-entropy centroids effectively capture the underly-



(a) From MHY to HMM. (b) From MHY to MHB.

Figure 6: Ground truth vs. predicted hateful ratios for one cluster per modality (1:visual, 2:textual, 3:audio). Smaller bar gaps indicate better intra-cluster prediction consistency.

ing core semantics and intents shared across domains, acting as the cores that bridge the target domain back to the source. Hence, our framework guides uncertain target samples toward these stable centroids for more robust adaptation.

4.5 Intra-cluster Diversity Evaluation

To evaluate the effectiveness of our intra-cluster diversity regularization, we compare the ground-truth and predicted hateful ratios within one cluster randomly selected from each modality across two target domains, as shown in Figure 6. The ground-truth class distributions within these clusters are relatively balanced, and our framework yields similarly balanced predictions, with class ratios consistently near 0.5. These results indicate that our method effectively mitigates the semantic collapse issue (cf. Figure 4) by preserving class diversity and preventing the source model from assigning all target samples to a single incorrect class.

4.6 Domain Distribution Visualization

To intuitively compare the capability of SCANNER and TENT in reducing domain discrepancy between source and target domains, we present UMAP visualizations (McInnes et al. 2018) of embedding distributions from the source (MHY) and adapted target (HMM) domains, as shown in Figure 7. Final representations are obtained by averaging outputs of modality-specific encoders. While TENT adapts via EM, the resulting source and target embeddings remain clearly separated, suggesting that implicit alignment fails to overcome substantial semantic shifts. In contrast, SCANNER achieves stronger alignment with considerable overlap between domains. This improvement stems from SCANNER 's use of invariant hateful cores as domain-bridging anchors that explicitly guide target samples toward source-aligned representations, effectively mitigating the influence of surface-level variations in evolving hateful content.

4.7 Case Study on Ambiguous Hateful Videos

We examine the adaptability of SCANNER on three representative target hateful videos, as illustrated in Figure 8. In all cases, both visual and textual content lack explicit hateful cues; the presented semantics are abstract and implicitly expressed, making it difficult to discern any hateful intent. Compared to the initial entropy produced by the

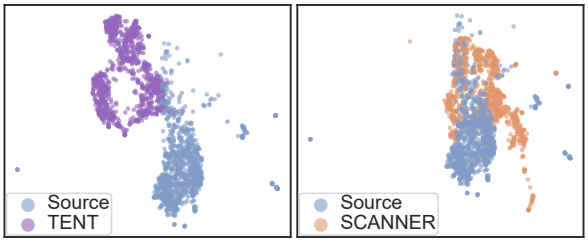


Figure 7: UMAP visualizations of embedding distributions for the source (MHY) and adapted target (HMM) domains on TENT and SCANNER.

	Case A	Case B	Case C
Videos			
Texts	2 gle SEQU THF CRY AGAIN.	Just enjoy it.	Copyright disclaimer under section 107...
Ground Truth	# Hateful	# Hateful	# Hateful
Source	0.201	0.038	0.107
TENT	0.560 # Non-hateful	0.032 # Non-hateful	0.074 # Non-hateful
SCANNER	0.006 # Hateful	0.005 # Hateful	0.041 # Hateful

Figure 8: Case study of SCANNER's adaptability on three randomly sampled hateful videos from the target domains. The numerical result is the prediction entropy.

source model, TENT exhibits limited adaptation capability, often leading to incorrect predictions and even performance degradation. In the particularly challenging Case A, TENT demonstrates high uncertainty toward subtly expressed hateful content, ultimately resulting in misclassifications. In contrast, SCANNER explicitly aligns target videos with the source domain via invariant hateful cores, yielding the lowest entropy values and correctly classifying all samples.

5 Conclusion

In this paper, we propose SCANNER, the first TTA framework tailored for HVD. We adopt a progressive expansion scheme, starting with a basic framework, dubbed CAN, which initially guides target samples toward the semantic centroids representing the invariant hateful cores. To further mitigate the impact of video samples distant from the semantic centroids, we extend CAN to SCAN by introducing sample-level adaptive weighting based on sample-to-centroid similarity. Finally, we incorporate an intra-cluster diversity loss into SCAN to avoid representation homogenization within cluster, thus forming the final framework, SCANNER. Extensive experiments on three benchmarks show that SCANNER outperforms all baselines in HVD.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62572097, No. 62176043, and No.U22A2097).

References

Cao, H.; Xu, Y.; Yang, J.; Yin, P.; Yuan, S.; and Xie, L. 2023a. Multi-Modal Continual Test-Time Adaptation for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18809–18819.

Cao, R.; Hee, M. S.; Kuek, A.; Chong, W.-H.; Lee, R. K.-W.; and Jiang, J. 2023b. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 5244–5252.

Das, M.; Raj, R.; Saha, P.; Mathew, B.; Gupta, M.; and Mukherjee, A. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 17, 1014–1023.

Dong, H.; Chatzi, E.; and Fink, O. 2025. Towards Robust Multimodal Open-set Test-time Adaptation via Adaptive Entropy-aware Optimization. In *International Conference on Learning Representations (ICLR)*.

Gan, Y.; Bai, Y.; Lou, Y.; Ma, X.; Zhang, R.; Shi, N.; and Luo, L. 2023. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 7595–7603.

Guo, Z.; and Jin, T. 2025. Smoothing the Shift: Towards Stable Test-Time Adaptation under Complex Multimodal Noises. In *International Conference on Learning Representations (ICLR)*.

Guo, Z.; Jin, T.; Xu, W.; Lin, W.; and Wu, Y. 2025. Bridging the gap for test-time multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 16987–16995.

Hamerly, G.; and Elkan, C. 2003. Learning the k in k-means. *Advances in Neural Information Processing systems (Neurips)*, 16.

Hoang, T. H.; Vo, M.; and Do, M. 2024. Persistent test-time adaptation in recurring testing scenarios. *Advances in Neural Information Processing systems (Neurips)*, 37: 123402–123442.

Hong, R.; Lang, J.; Zhong, T.; and Zhou, F. 2025. Borrowing Eyes for the Blind Spot: Overcoming Data Scarcity in Malicious Video Detection via Cross-Domain Retrieval Augmentation. In *IEEE International Conference on Computer Vision (ICCV)*.

Koushik, G. A.; Kanojia, D.; and Treharne, H. 2025. Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2014–2023.

Lang, J.; Hong, R.; Xu, J.; Li, Y.; Xu, X.; and Zhou, F. 2025. Biting Off More Than You Can Detect: Retrieval-Augmented Multimodal Experts for Short Video Hate Detection. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2763–2774.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, 6028–6039. PMLR.

Liu, J.; Yang, S.; Jia, P.; Lu, M.; Guo, Y.; Xue, W.; and Zhang, S. 2024. ViDA: Homeostatic Visual Domain Adapter for Continual Test Time Adaptation. *International Conference on Learning Representations (ICLR)*.

Ma, J. 2024. Improved self-training for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23701–23710.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Mei, J.; Chen, J.; Yang, G.; Lin, W.; and Byrne, B. 2025. Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection. *arXiv preprint arXiv:2502.13061*.

Nan, H.; Wang, Z.; Zhang, Y.; Zhang, X.; and Zeng, S. 2025. ResponSight: Explainable and Collaborative Moderation Approach for Responsible Video Content in UGC Platform. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–12.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *International Conference on Learning Representations (ICLR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763. PmLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing systems (Neurips)*, 33: 11539–11551.

Tsai, Y.-Y.; Chen, F.-C.; Chen, A. Y. C.; Yang, J.; Su, C.-C.; Sun, M.; and Kuo, C.-H. 2024. GDA: Generalized Diffusion for Robust Test-time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23242–23251.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing systems (Neurips)*, 30.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations (ICLR)*.

Wang, H.; Yang, T. R.; Naseem, U.; and Lee, R. K.-W. 2024a. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 7493–7502.

Wang, M.; Liu, Y.; Yuan, J.; Wang, S.; Wang, Z.; and Wang, W. 2024b. Inter-class and inter-domain semantic augmentation for domain generalization. *IEEE Transactions on Image Processing (TIP)*, 33: 1338–1347.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7201–7211.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations (ICLR)*.

Zhang, L.; Jin, L.; Sun, X.; Xu, G.; Zhang, Z.; Li, X.; Liu, N.; Liu, Q.; and Yan, S. 2023. TOT: topology-aware optimal transport for multimodal hate detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 4884–4892.

Zhou, X.; Tian, Z.; Zhang, B.; Zhang, Y.; Cheung, K. C.; See, S.; Yang, H.; Zhou, Y.; and Zhang, N. L. 2025. Test-time adaptation on noisy data via model-pruning-based filtering and flatness-aware entropy minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 10852–10860.

Zhou, Z.; Guo, L.-Z.; Jia, L.-H.; Zhang, D.; and Li, Y.-F. 2023. Ods: Test-time adaptation in the presence of open-world data shift. In *International Conference on Machine Learning (ICML)*, 42574–42588. PMLR.