

# Multimodal Clickbait Detection by De-confounding Biases Using Causal Representation Inference

Jianxing Yu<sup>\*</sup>, Shiqi Wang<sup>†</sup>, Han Yin<sup>†</sup>, Zhenlong Sun, Ruobing Xie, Bo Zhang, Yanghui Rao

School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, 519082, China

International Campus, Zhejiang University, Haining, 314400, China

Key Laboratory of Sustainable Tourism Smart Assessment Technology, Ministry of Culture and Tourism

WeChat Search Application Department, Tencent, Beijing, 100080

Pazhou Lab, Guangzhou, 510330, China

{yujx26, wangshq25, raoyangh}@mail.sysu.edu.cn, han.22@intl.zju.edu.cn

richardsun@tencent.com, xrbsnowing@163.com, nevinzhang@tencent.com

## Abstract

This paper focuses on detecting clickbait posts on the Web. These posts often use eye-catching disinformation in mixed modalities to mislead users to click for profit. That affects the user experience and thus would be blocked by content provider. To escape detection, malicious creators use tricks to add some irrelevant non-bait content into bait posts, dressing them up as legal to fool the detector. This content often has biased relations with non-bait labels, yet traditional detectors tend to make predictions based on simple co-occurrence rather than grasping inherent factors that lead to malicious behavior. This spurious bias would easily cause misjudgments. To address this problem, we propose a new debiased method based on causal inference. We first employ a set of features in multiple modalities to characterize the posts. Considering these features are often mixed up with unknown biases, we then disentangle three kinds of latent factors from them, including the invariant factor that indicates intrinsic bait intention; the causal factor which reflects deceptive patterns in a certain scenario, and non-causal noise. By eliminating the noise that causes bias, we can use invariant and causal factors to build a robust model with good generalization ability. Experiments on three popular datasets show the effectiveness of our approach.

## 1 Introduction

With the rapid development of social media, people share views and advertise products by posting content on platforms (Liao et al., 2021) like WeChat, Twitter, Instagram, etc. To increase viewership and obtain more advertising revenue, unscrupulous creators misuse these platforms to publish deceptive, poor-quality posts. Such posts are often described with a catchy thumbnail or a sensational headline. For example, the posts have a sexy thumbnail, with

curiosity-inciting phrases like “*You Won’t Believe*”, and “*X Reasons Why*” in their headlines. That would bait readers to click on the linked articles. However, these articles would be unrelated to the thumbnails and headlines of the posts. Moreover, they are often full of hoaxes, rumors, and fake news, which not only degrade users’ experience but also affect the credibility of the platforms (Zhu et al., 2023). Thus, the detection of such clickbait posts has great commercial value for social media.

Due to the large volume of emerging posts, a manual review of the clickbait is infeasible. Machine detection has become a hot topic (Comito et al., 2023). The detective sources can be summarized into two categories (Yadav and Bansal, 2023). The first is based on social behavior. Since bait posts are required to spread in social networks to expand their influence, typical propagation characteristics can be observed. It can be detected based on social metadata like comments, the number of views, shares, likes, etc (Agarwal et al., 2023). That requires a rich collection of user feedback for judgment, but this feedback is often delayed or some users do not even share it. As a result, malicious posts can only be found after they have been widely spread, which is too late for online applications. Another direction is to analyze the post contents. The bait posts often have certain linguistic characteristics in terms of deceptive words, syntax, subjectivity, writing style (Zhou et al., 2020), and even punctuation (Coste and Bufnea, 2021). Besides, there may be abnormal relations between their subparts in various modalities, such as a rumor article text with visual thumbnails of an irrelevant actor to attract clicks. Early works design a set of rules to detect these features and relations, but the rules are hand-crafted and non-scalable. Current mainstream methods turn to the neural model for improving scalability. They seek correlations between the post content and labels to make predictions.

However, to escape detection, malicious creators

<sup>\*</sup>Corresponding author.

<sup>†</sup>These authors have contributed equally to this work.

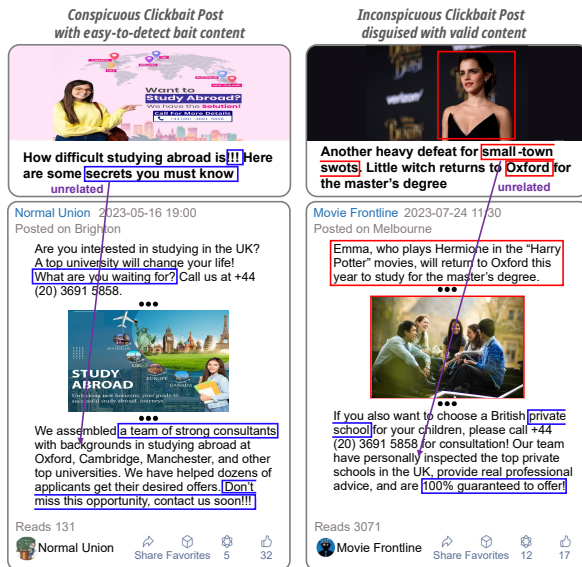


Figure 1: Clickbait samples. The purple arrow indicates inconsistencies between the headline and its linked article. The simple clickbait post contains conspicuous bait-indicative words or advertising content (marked with blue boxes) which is easily detected. The complex one disguises the bait content with some valid content (in red boxes) and makes it look inconspicuous, thus deceiving and escaping the detector.

with use some unrelated content to disguise a bait post as a valid one. As shown in Fig.(1), in the left post, some of the conspicuous bait thumbnail and text are replaced with unarmful content, making it look like a legal one on the right. Such content may co-occur with non-bait label usually, creating spurious correlations, i.e., the unstable and confounding patterns in the feature space (Dogra et al., 2022). That would easily cheat the model to fail to discover the hidden bait. Respectively, some valid posts may be misjudged as bait simply because they contain content that co-occurs with a bait label. Such biased correlations lead to high misjudgments. The bias is inadequate to differentiate by a naive encoder, which is designed for generic content understanding tasks such as text classification. The encoder focuses on general embedded representations rather than key factors that cause fraudulent behaviors (Mukherjee et al., 2022). These behaviors are not static, their topics and types may change over time. The writing styles would also vary for different authors, but such disguise tricks commonly follow some styles and patterns in a particular scenario. On the other hand, the data-driven neural model relies on scarce labeled data heavily. To derive a robust model, we

have to collect data on all scenarios for supervised training. Such uniform data is either unavailable or costly to acquire. As a result, the statistics on limited samples are not significant, which is easy to train the model to remember some inessential spurious correlations. That would harm the model’s robustness. A simple solution is to customize a representation that perceives posts’ quality, but that would face tedious hand-crafted engineering.

Motivated by above observations, we propose a new debiased approach to detect clickbait posts. In detail, we first represent the given posts based on a set of multimodal features, including textual and visual features, linguistic and cross-modal features, as well as features of the creators’ profiles. Considering these mixed representations contain unknown biases, we resort to causal representation learning, which is good at eliminating non-causal spurious noise. The representations are disentangled into three key latent factors, including (1) invariant factor that indicates the inherent bait’s intention and post quality; (2) causal factor in a specific scenario; and (3) non-causal noise factor. The noise like unrelated words and topics is defined under a specific scenario, and each is unique. The invariant factor is obtained by invariant risk minimization, scenarios are estimated from the data, and the causal factor is learned by contrastive learning. By removing the noise and using the remaining invariant and causal factors, we can build a robust clickbait detector. It can combat spurious bias and generalize well on newly formed bait subspecies. To facilitate training, we further develop a data augmentation technique to alleviate the labeled data scarcity problem. Extensive experiments on three real-world datasets show the effectiveness of our approach.

The main contributions of this paper include,

- We reveal the issue of bait subspecies evolution via disguise and point out the challenges of the resulting spurious bias in the field of multimodal clickbait detection, which is new.
- We propose a new debiased model from a view of causal inference. It explores a prior causal structure to elicit latent key factors that reflect posts’ quality. That can alleviate spurious bias and achieve better generalization ability.
- We conduct extensive experiments to fully evaluate the effectiveness of our method.

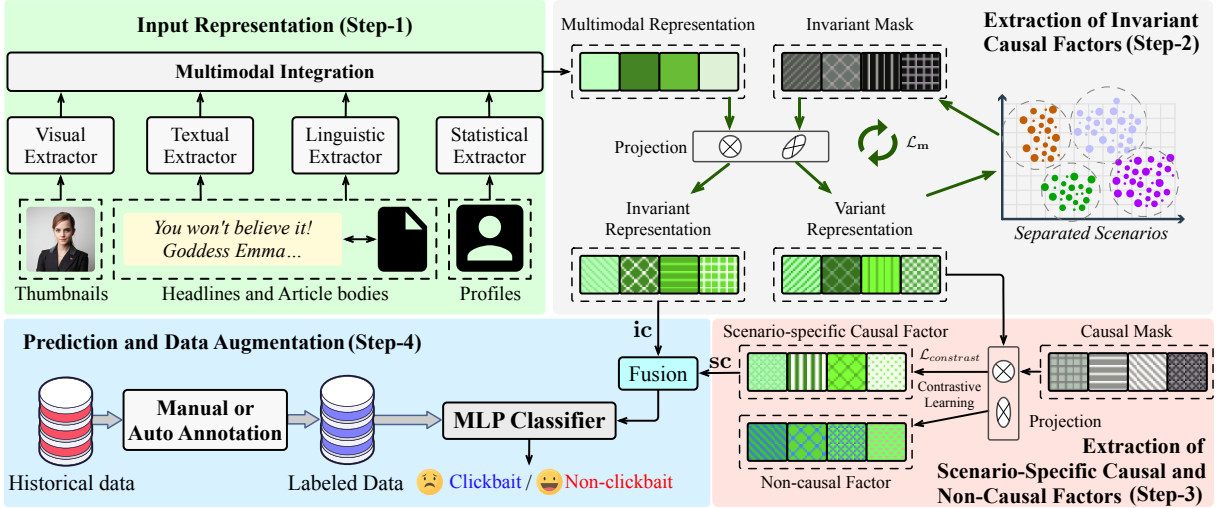


Figure 2: The overview framework of our causal clickbait detector.

## 2 Approach

Fig.(2) shows our framework with four steps. We first extract multimodal features from the posts. By causal inference, we then disentangle the invariant factors from them, and separate the remaining parts into the causal factor and non-causal factor. Finally, we make the prediction based on the invariant and causal factors. Next, we define some notations and elaborate on each component of our approach.

### 2.1 Notations and Problem Formulation

**Clickbait detection.** Given a post  $x_i$ , our task aims to learn a model  $\mathcal{F}(y_i|x_i, \Theta_{\mathcal{F}})$  to predict whether it is clickbait ( $y_i = 1$ ), where  $\Theta_{\mathcal{F}}$  is the model’s parameter set,  $x_i$  denotes the post contents, including the textual headline, visual thumbnail and their linked mixed-modal article. These sub-parts have some deceptive characteristics and relations, such as containing malicious rhetoric like exaggeration, eroticism, bluffing, weirdness, and distortion; the headline or thumbnail is seductive but unrelated to the article. In addition, the creators constantly yield new bait subspecies to avoid being detected and seized. Thus, our target is to find an optimized  $\hat{\Theta}_{\mathcal{F}}$  by  $\arg \min_{\hat{\Theta}_{\mathcal{F}}} \mathcal{L}(\hat{\mathcal{F}}(y_i|x_i, \hat{\Theta}_{\mathcal{F}})|\mathcal{D}^{tr})$  which can be generalized well to the test set  $\mathcal{D}^{te}$  with a lowest cost  $\mathcal{L}(\mathcal{D}^{te})$ , where  $\mathcal{D}^{tr}$  is the training set,  $\mathcal{L}(\cdot)$  denotes a cross-entropy classification loss.

**Eliminating spurious correlations.** As displayed in Fig.3.(a), traditional methods often make predictions based on the co-occurrence between post features  $X$  and labels  $Y$ . Ideally,  $X$  is expected to reflect the hidden malicious intention of the post. However, it comes from generic encoders without

considering the fraud behavior. That inevitably introduces some false correlations, such as wrongly linking some trivial but irrelevant terms or images with non-bait label. Besides, some bait modes are stable while some writing styles would change in various scenarios, such as periods, types, and creators. It is easy to make erroneous predictions without grasping this distinction. A simple solution is to dissociate  $X$  into several factors, i.e., an invariant factor ( $IC$ ) that reflects malicious intention, a causal factor for a certain scenario ( $SC$ ), and non-causal noise ( $NF$ ); and then eliminate the effect of noise  $NF$  on  $Y$ , as Fig.3.(b). However, without necessary constraints, the spurious correlations in  $NF$  may permeate via  $IC$  and  $SC$  to harm the prediction of  $Y$ . To prevent this effect, we introduce a causal structure to regularize these latent variables, as Fig.3.(c). We first use a confounder  $C$  to capture mixed relations of three factors under the condition of a certain scenario  $S$ . We then isolate the invariant  $IC$  by blocking the influence of  $C$  on  $IC$ . The bias in  $C$  cannot affect  $Y$  through  $IC$ . To fully put away the noise  $NF$ , we cut the backdoor paths of  $C$  and  $S$  on  $NF$ . The  $S$  and  $C$  can only impact  $Y$  via  $SC$ . In this way, we can independently elicit salient unbiased factors without mutual influence, which can generalize well to new bait subspecies.

### 2.2 Multimodal Feature Extraction

To fully capture the characteristics of the posts, we extract five kinds of features in multiple modalities. More details are exhibited in the Appendix A.

**(1) Visual Features:** We encode each post image (e.g. thumbnail and article figure) by transformer-

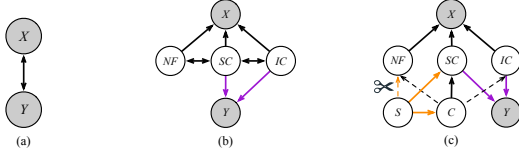


Figure 3: Causal structure for de-confounding biases. Gray and white nodes represent observable and unobserved variables, respectively; the bidirectional and unidirectional arrows denote correlations and causalities, respectively; purple arrows are key causalities determining result  $Y$ ; and orange arrows refer to scenario effects.

based *Swin-T* (Liu et al., 2021). To grasp the fine-grained features in the image, we conduct facial recognition and object detection by *DNN* and *RetinaNet* models (Lin et al., 2017), respectively.

**(2) Textual Features:** For the text in the post, such as headline, description, and associated article, we first remove hashtag, mention, punctuation, and URL. We then tokenize each and embed it by  $BERT_{base}$  pre-trained model (Devlin et al., 2019). Since clickbait posts often contain seductive text in the thumbnails, we utilize the *OCR* technique (Li et al., 2023) to extract the text and encode it.

**(3) Cross-modal Features:** The clickbait posts often contain inconsistencies, such as using inviting thumbnails to lure clicking the unrelated article. Verifying this cross-modal mismatch between the post’s sub-parts can help prediction. Due to the heterogeneous gap, it is hard to use this inter-modal complementary benefit directly. To tackle this issue, we employ a *CT Transformer* (Lu et al., 2019), which is good at capturing multi-modal relations.

**(4) Linguistic Features:** To capture the bait patterns and writing styles, we extract six kinds of features based on thumbnail, headline, and article body, including article body-thumbnail disparity, article body-headline disparity, thumbnail-headline disparity, sentiment of headline, lexical analysis of headline and baitiness analysis of headline.

**(5) Profile Features:** The posts’ quality often depends on the creators, i.e., bad creators tend to write malicious posts. To characterize the quality of each creator  $u_j$ , we extract features based on its profiles, involving the register age, self-description and screen name of  $u_j$ , number of followers for  $u_j$ , number of users that  $u_j$  is following, number of created posts of  $u_j$ , time elapsed after  $u_j$ ’s first post, whether the  $u_j$  account is verified or not, whether  $u_j$  allows the geo-spatial positioning, time difference between the source post time and  $u_j$ ’s share time, and length of retweet path between  $u_j$  and a

source post. The length is 1 if  $u_j$  retweets the post. By concatenating all these features, we can obtain a  $d$ -dimensional vector  $\mathbf{x}_i$  as a post’s representation.

### 2.3 Disentanglement of Invariant Factors

The generic  $\mathbf{x}_i$  may contain spurious correlations, which are mostly unreliable across various scenarios. For example, the false correlation between an unrelated actress and the bait label will change in different contexts of posts. In contrast, it is usually stable for the bait behavior which is caused by some common factors, such as underlying language patterns and deceptive habits. By seeking commonalities in various scenarios, we can capture these inherent factors to achieve better robustness.

**Casual Inference:** We use an invariance mask  $\mathbf{m} \in \mathbb{R}^d$  to dissociate the invariant characteristics of the original representation  $\mathbf{x}_i$  as  $\mathbf{ic}_i = \mathbf{m} \odot \mathbf{x}_i$ , where  $\odot$  is the element-wise product operator. The opposite of  $\mathbf{m}$  is the variance mask, which can extract the variant representation  $\mathbf{vc}_i = (1 - \mathbf{m}) \odot \mathbf{x}_i$ . Ideally,  $\mathbf{ic}_i$  should have identical joint distributions with the unbiased variable across various scenarios. We pursue it by optimizing  $\mathbf{m}$  with an invariant risk minimization (*IRM*) (Arjovsky et al., 2019) objective. *IRM* introduces a penalty for the variation of empirical risks in all scenarios. It tries to mask some features and calculate their impact on the results, so as to find out factors that have a stable and significant impact on the results. That encourages the mask to suppress spurious features and emphasize the causally invariant ones. We formulate *IRM* loss as a gradient norm penalty over the empirical risk  $\mathcal{L}_s$  in each training scenario as Eq.(1):

$$\mathcal{L}_{\mathbf{m}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} [\mathcal{L}_s(\mathbf{m}, \mathcal{F}_{\mathbf{m}}) + \alpha \|\nabla_{\mathcal{F}_{\mathbf{m}}} \mathcal{L}_s(\mathbf{m}, \mathcal{F}_{\mathbf{m}})\|^2] + \beta \|\mathbf{m}\|^2. \quad (1)$$

where  $\alpha$  and  $\beta$  are trade-off factors, the second term is the constraint over the variance of scenarios, and the third one is a regularization term.  $\mathcal{L}_s = \mathcal{L}(\mathcal{F}_{\mathbf{m}}(y_i|x_i, \Theta_{\mathcal{F}_{\mathbf{m}}})|D_s^{tr})$  is the classified loss on the training subset  $D_s^{tr}$  under the specific scenario  $s$ . This objective can enforce the mask  $\mathbf{m}$  to seek stable inherent patterns in the context, instead of learning an average effect of spurious correlations.

**Scenario Estimation:** The key to optimizing  $\mathbf{m}$  lies in the division of scenarios. We thus utilize scenarios to infer the spurious correlations and determine each with a set of operations. In each scenario, the posts have certain characteristics, such as

writing styles, hot topics, and evolving bait patterns. Since the scenario is unprovided, we propose to estimate it from the data with two iterative phases. First, we learn a scenario predictive model to fit the data distribution. We then reassign the samples into appropriate scenario subsets. To initialize this iterative process, we randomly assign training samples to each scenario subset  $\mathcal{D}_s^{tr}$ . By alternatively optimizing the data-fit *IRM* objective and the estimated scenarios, finally we can learn the invariant factors across scenarios.

(1) *Scenario Division*. We observe that the key difference among various scenarios lie in their characteristics of spurious correlations. We thus explore such correlations to learn a scenario predictive model  $\Phi_s$  for each  $s \in \mathcal{S}$ . In detail, for the  $i^{th}$  sample,  $\Phi_s$  evaluates the likelihood that it belongs to the scenario  $s$  based on its variant representation  $\mathbf{vc}_i$ . Parameterized by  $\theta_s$ , the model  $\Phi_s$  aims to minimize the prediction loss on the training subset  $\mathcal{D}_s^{tr}$ , as Eq.(2). Relying on this objective,  $\Phi_s$  can be learned without interference from other scenario data, which can better handle the spurious bias.

$$\min_{\theta_s} \mathcal{L}(\Phi_s(\mathbf{vc}_i|\theta_s)|\mathcal{D}_s^{tr}). \quad (2)$$

(2) *Samples Reallocation*. Based on the estimation in phase 1, we obtain  $|\mathcal{S}|$  scenario models which indicate different types of spurious biases. Next, we reallocate all samples to the appropriate scenarios according to their spurious correlations. In detail, we feed the variant representation  $\mathbf{vc}_i$  into each scenario model  $\Phi_s$ . Each sample is then assigned to a scenario with the highest likelihood as Eq.(3).

$$s(i) \leftarrow \arg \max_{s \in \mathcal{S}} \Phi_s(\mathbf{vc}_i|\theta_s). \quad (3)$$

By running these two phases until convergence, we obtain stable scenario subsets  $\{\mathcal{D}_s^{tr}|s \in \mathcal{S}\}$ . In turn, we can optimize the invariance mask  $\mathbf{m}$ . Finally, we learn an optimal  $\mathbf{m}$  to derive a universal causal factor that is invariant in most situations.

## 2.4 Dissociation of Causal Factor from Noise

After separating the invariant factor, the remaining part  $\mathbf{vc}_i$  is a mixture of causality and noise in a specific scenario. We thus further elicit the valuable scenario-specific causal factor  $\mathbf{sc}_i$  from  $\mathbf{vc}_i$ . We first project it into a latent embedding space, as  $\xi(\mathbf{vc}_i)$ .  $\xi(\cdot)$  is a network with a multi-layer perceptron, which is parameterized by  $\theta_\xi$ . When designing  $\xi$ , we do not inject any scenario category information. That allows us to adapt to some new

scenarios without refactoring  $\xi$  or relearning the parameters. We then output a causal perception mask  $\gamma = \text{Gumbel} - \text{SoftMax}(\xi(\mathbf{vc}_i), kd)$  using the *Gumbel-SoftMax* technique. The mask  $\gamma$  sets the values of some dimension to 0 and retains  $kd$  effective dimensions. By using the mask  $\gamma$ , we can extract the causal representation as  $\mathbf{sc}_i = \gamma \odot \mathbf{vc}_i$ .

To facilitate the separation of causal factors from noises, we introduce contrastive constraints based on the causal interventions. After the causal vector  $\mathbf{sc}_i$  is extracted from  $\mathbf{vc}_i$ , we collect the remaining non-causal one  $\mathbf{nf}_i = (1 - \gamma) \odot \mathbf{vc}_i$ .  $\mathbf{nf}_i$  can be used as a contrastive learning signal with the loss as Eq.(4). For a clickbait sample ( $y = 1$ ), its non-causal feature  $\mathbf{nf}_i$  exactly blocks out the identifiable clickbait characteristics. When replacing the scenario-specific causal representation  $\mathbf{sc}_i$  with  $\mathbf{nf}_i$ , the prediction is supposed to be the opposite, i.e., as non-clickbait. Accordingly, for the non-clickbait sample ( $y = 0$ ), since  $\mathbf{sc}_i$  does not indicate any clickbait characteristics, using  $\mathbf{nf}_i$  for detection would not change the prediction.

$$\mathcal{L}_{contrastive} = \mathcal{L}(\mathcal{F}_\xi(0|\mathbf{nf}_i, \Theta_{\mathcal{F}_\xi})|\mathcal{D}^{tr}). \quad (4)$$

## 2.5 Prediction and Data Augmentation

By repeating the running flow (i.e., vector masking  $\rightarrow$  scenario division  $\rightarrow$  mask learning) for  $T$  times till converged, we can obtain the key causal representations, namely, the invariant causal factor and scenario-specific causal factor. We concatenate these two vectors to train a *Multilayer Perceptron* classifier, with an objective as Eq.(5).

$$\arg \min_{\hat{\Theta}_{\mathcal{F}}} \mathcal{L}(\hat{\mathcal{F}}(y_i|[\mathbf{ic}_i; \mathbf{sc}_i], \hat{\Theta}_{\mathcal{F}})|\mathcal{D}^{tr}). \quad (5)$$

To better train the model, we further employ data augmentation to collect pseudo-labeled data. Since clickbait posts need to be widely spread to gain benefits, social behavior is often an effective clue. We thus design heuristic rules based on social metadata, such as share frequency, and viewing time. This metadata may be lacking in new cases, but it is sufficient in historical data. That can provide abundant clickbait cases as training data to reduce labeled costs. In this way, our model can work well even if only limited data is provided in some applications. More details are shown in Appendix D.4.

## 3 Evaluations

We fully conducted experiments with qualitative and quantitative analyses to evaluate our approach.

### 3.1 Data and Experimental Settings

We performed evaluations on three popular real-world datasets, including *CLDInst* (Ha et al., 2018), *Clickbait17* (Potthast et al., 2018) and *FakeNewsNet* (Shu et al., 2020a). By crowd-sourcing, these datasets were split as bait/non-bait sets with the size of 4k/3k, 9k/29k, and 5k/17k posts, respectively. For each sample, we crawled the original post from its *URL* to collect multimodal data, such as the thumbnail and creator’s profile.

- *CLDInst* covers 7,769 fashion-related posts crawled from *Instagram*. They are judged by annotators employed from crowdsourcing websites. A total of 4,260 posts are tagged as clickbait.

- *Clickbait17* contains 38,517 Twitter posts as well as their linked articles from 27 major US news publishers. To avoid bias, a maximum of 10 posts per day was sampled for each publisher, with 9,276 posts tagging as clickbait.

- *FakeNewsNet* is a large-scale multimodal news dataset that contains over 23k articles with tagged fake/real labels from the websites of *PolitiFact* and *GossipCop*. It has a rich social context, with 432 fake and 624 real samples from *PolitiFact*, as well as 5,323 fake and 16,817 real cases from *GossipCop*. Similar to bait posts, fake samples often suffer from issues like irrelevance, inconsistency, etc. This dataset can be used to evaluate the model’s ability to recognize bait-like low-quality content.

Each dataset was split into train/validation/test sets. We tuned the model on a validation set and reported results on the test set. In addition, we further evaluated the value of data augmentation technique in Appendix D.4. We employed four typical metrics in the field of classification for evaluations, including accuracy (*ACC*), precision (*PRE*), recall (*REC*), and F1-score (*F1*). To tackle the class imbalance problem, we trained all evaluated methods by using the oversampling technique. To reduce bias, we repeated running 20 times and reported the average performance. In addition, the configurations of all evaluated methods were shown in Appendix B.

### 3.2 Comparisons against State-of-the-arts

To verify the effectiveness of our method, we compared it against six typical baselines in the field of clickbait detection. including (1) *dEFEND* (Shu et al., 2019a), a co-attention-based model that predicted based on both post content and user profiles; (2) *HPFN* (Shu et al., 2020b), which made

judgments based on posts’ propagation on social network; (3) *MCAN* (Wu et al., 2021), a multimodal model that captured both textual and visual features by stacked co-attention layers; (4) *CPDM* (Mowar et al., 2021), using ensemble classifier to perceive inconsistencies among content, headline and thumbnail; (5) *CCD* (Chen et al., 2023), a model based on causal intervention and counterfactual reasoning; (6) *VLP* (Wang et al., 2023), a multimodal pre-trained detector.

As displayed in Table 1, our method achieved the best performance. The outperformance was over the best baselines (e.g., *VLP*) on *CLDInst*, *Clickbait17*, and *FakeNewsNet* by 3.78%, 4.66%, and 4.02% in terms of the accuracy metric, respectively. Methods with multimodal features (i.e., *MCAN*, *CPDM*, *CCD* and *VLP*) showed better performance, since these features provided useful discriminant clues. In addition, our method performed better than the causal baseline *CCD*. *CCD* only tackled the misalignment among the textual and visual features but neglected the spurious bias and disguised content that are widespread in bait posts. Besides, we observed on the larger datasets e.g., *Clickbait17* and *FakeNewsNet*, our outperformance was bigger. The reason may be that spurious bias in these datasets was more extensive, and our debiased gain was greater. To evaluate it, we further selected 500 test samples randomly from each dataset and annotated each post manually. We found that approximately 23%, 26%, 27% of the posts were disguised type, respectively. The datasets *Clickbait17* and *FakeNewsNet* were more complicated, having a larger percentage of disguised samples. Moreover, we provided the precision-recall curves on three datasets in Fig.(4). Our model achieved the best precision across all recall levels. That reflected the effectiveness of our approach in eliminating spurious bias. Ours can obtain a high recall rate and well identify the bait posts dressed up as valid ones.

### 3.3 Ablation Studies

To gain insight into the relative contributions of each component in our approach, we performed ablation studies on four aspects, including (1) *w/o MF* that discarded the multimodal representation module and relied solely on text encoders; (2) *w/o EICF* that dropped the invariant causal factor by removing the *IRM* regularization in Eq.(1); (3) *w/o ENF* which threw away the scenario learning module, randomly assigned scenarios to samples; (4) *w/o ESCF* that removed scenario-specific causal

Methods	CLDInst				Clickbait17				FakeNewsNet			
	ACC $\uparrow$	PRE $\uparrow$	REC $\uparrow$	F1 $\uparrow$	ACC $\uparrow$	PRE $\uparrow$	REC $\uparrow$	F1 $\uparrow$	ACC $\uparrow$	PRE $\uparrow$	REC $\uparrow$	F1 $\uparrow$
<b>dEFEND</b>	76.58 $\pm$ 0.06	74.03 $\pm$ 0.12	76.46 $\pm$ 0.29	75.23 $\pm$ 0.37	82.15 $\pm$ 0.33	76.74 $\pm$ 0.08	79.22 $\pm$ 0.27	77.96 $\pm$ 0.10	81.06 $\pm$ 0.21	74.21 $\pm$ 0.17	78.74 $\pm$ 0.36	76.41 $\pm$ 0.28
<b>HPFN</b>	73.15 $\pm$ 0.14	72.12 $\pm$ 0.31	70.23 $\pm$ 0.27	71.16 $\pm$ 0.15	81.68 $\pm$ 0.26	84.25 $\pm$ 0.13	82.35 $\pm$ 0.34	83.29 $\pm$ 0.22	84.82 $\pm$ 0.17	84.43 $\pm$ 0.13	85.68 $\pm$ 0.24	85.05 $\pm$ 0.23
<b>MCAN</b>	79.21 $\pm$ 0.18	77.45 $\pm$ 0.07	77.80 $\pm$ 0.19	77.62 $\pm$ 0.28	84.51 $\pm$ 0.20	85.56 $\pm$ 0.09	82.74 $\pm$ 0.33	84.13 $\pm$ 0.14	83.82 $\pm$ 0.29	85.13 $\pm$ 0.37	82.21 $\pm$ 0.18	83.64 $\pm$ 0.15
<b>CPDM</b>	80.04 $\pm$ 0.23	77.89 $\pm$ 0.42	79.31 $\pm$ 0.11	78.59 $\pm$ 0.39	86.14 $\pm$ 0.07	86.30 $\pm$ 0.10	83.07 $\pm$ 0.21	84.65 $\pm$ 0.28	84.27 $\pm$ 0.13	85.18 $\pm$ 0.22	82.44 $\pm$ 0.14	83.79 $\pm$ 0.31
<b>CCD</b>	82.77 $\pm$ 0.14	83.13 $\pm$ 0.10	82.91 $\pm$ 0.16	83.02 $\pm$ 0.24	88.36 $\pm$ 0.22	87.61 $\pm$ 0.15	86.46 $\pm$ 0.09	87.03 $\pm$ 0.27	87.72 $\pm$ 0.13	85.96 $\pm$ 0.20	86.46 $\pm$ 0.15	86.21 $\pm$ 0.14
<b>VLP</b>	85.56 $\pm$ 0.16	84.25 $\pm$ 0.07	83.87 $\pm$ 0.20	84.06 $\pm$ 0.12	88.70 $\pm$ 0.26	87.34 $\pm$ 0.37	86.02 $\pm$ 0.41	86.67 $\pm$ 0.24	88.02 $\pm$ 0.10	87.25 $\pm$ 0.16	86.23 $\pm$ 0.19	86.74 $\pm$ 0.30
<b>Ours</b>	88.79 $\pm$ 0.11	88.73 $\pm$ 0.04	89.16 $\pm$ 0.32	88.94 $\pm$ 0.14	92.83 $\pm$ 0.27	93.45 $\pm$ 0.17	93.59 $\pm$ 0.15	93.52 $\pm$ 0.20	91.56 $\pm$ 0.07	92.74 $\pm$ 0.38	92.97 $\pm$ 0.13	92.85 $\pm$ 0.08

Table 1: Comparisons of all methods. The improvements were significant using a statistic t-test with p-value<0.005.

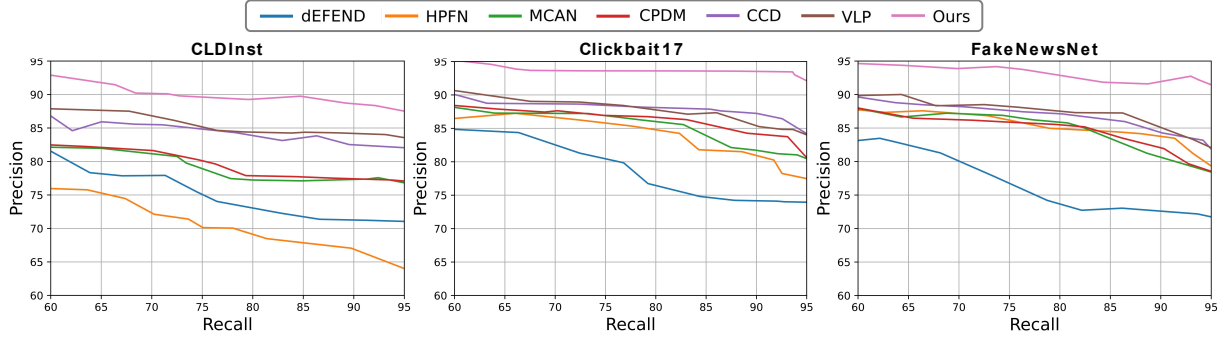


Figure 4: PR curves of all models on three datasets.

factor by deleting the module in Eq.(4).

As shown in Table 2, the ablation on all evaluated components led to a significant performance drop. This reflected that all four modules were indispensable. Among them, discarding the *EICF* module caused the most substantial decrease, i.e., more than 6.33% and 6.77% drops in terms of the accuracy and F1 metrics, respectively. That indicated the usefulness of eliminating spurious bias, which can improve discriminant and generalization ability. The multimodal module was also impactful, with its removal leading to a reduction of around 4.61%  $\sim$  5.63% in terms of accuracy. This demonstrated the benefits of inter-modal signals to overcome unimodal one-sidedness. In addition, the *ENF* and *ESCF* modules focus on scenario-specific and non-causal factors, which also led to noticeable performance drops when they were ablated. The results validated the rationality of our design. Ablation studies on other metrics were shown in Appendix D.5.

Datasets	CLDInst		Clickbait17		FakeNewsNet	
	ACC $\downarrow$	F1 $\downarrow$	ACC $\downarrow$	F1 $\downarrow$	ACC $\downarrow$	F1 $\downarrow$
w/o MF	-4.09 $\pm$ 0.10	-4.72 $\pm$ 0.08	-5.23 $\pm$ 0.13	-5.96 $\pm$ 0.06	-4.81 $\pm$ 0.09	-5.60 $\pm$ 0.07
w/o EICF	<b>-5.62<math>\pm</math>0.14</b>	<b>-6.03<math>\pm</math>0.06</b>	<b>-7.22<math>\pm</math>0.09</b>	<b>-8.06<math>\pm</math>0.12</b>	<b>-6.62<math>\pm</math>0.18</b>	<b>-7.07<math>\pm</math>0.09</b>
w/o ESCF	-2.88 $\pm$ 0.05	-3.70 $\pm$ 0.08	-3.86 $\pm$ 0.16	-4.27 $\pm$ 0.11	-3.64 $\pm$ 0.03	-3.87 $\pm$ 0.14
w/o ENF	-3.33 $\pm$ 0.06	-3.92 $\pm$ 0.10	-4.11 $\pm$ 0.18	-4.73 $\pm$ 0.09	-3.77 $\pm$ 0.15	-4.31 $\pm$ 0.05

Table 2: Ablation study with t-test, p-value<0.005.

### 3.4 Study of the Mask Mechanism

Relying on the mask vector  $\mathbf{m}$ , our model can extract a causal invariant factor applicable to most scenarios. We observed that the setting of  $\mathbf{m}$  can be binary ( $B$ ) or float ( $F$ ), and it can be integrated into the objective Eq.(1) by the  $L_2$  norm or a  $L_0$  regularizer. To better understand how the mask  $\mathbf{m}$  impacted the performance, we tested four  $\mathbf{m}$  configurations: ( $B + L_0$ ), ( $B + L_2$ ), ( $F + L_0$ ) and ( $F + L_2$ ). As shown in Fig.(5), we found that the float masks ( $F$ ) consistently outperformed binary masks ( $B$ ). This verified that keeping the continuous mask values provided more representational power than binary ones. In addition,  $L_2$  regularization worked better than  $L_0$ . Among all datasets,  $L_2$  norm outperformed  $L_0$  norm by around 1.23%  $\sim$  2.01% on the F1 score. The sparse  $L_0$  regularization might suppress informative dimensions, while  $L_2$  allowed more flexibility. More evaluations on the mask mechanism were shown in Appendix D.1.

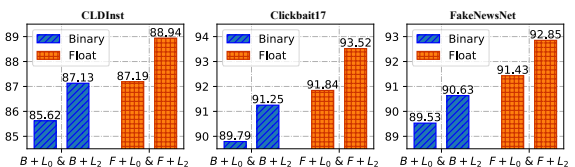


Figure 5: Evaluation the impact of  $\mathbf{m}$  settings on F1.

### 3.5 Study of the Scenario

Scenario is valuable to learn invariant representations. These stable features can effectively improve the robustness of the model and avoid being deceived by disguised content. To study the characteristics of scenarios, we first checked their separability. It was estimated from the data based on alternating optimization. That is, the samples were grouped to form new scenarios; in turn, after the scenarios were updated, the samples were moved to partition again. If most samples no longer changed, the scenarios can be viewed to be divisible. Thus, we calculated the ratio of moved samples to infer the stability of scenarios. If a sample cannot fit the current scenario and needed to be reassigned to a new one, we counted it as a moved case. As shown in Fig.(6), the curves of the moved rate on three datasets converged after being repeated around 10 rounds. That indicated the scenarios can be separated and help to capture the spurious bias well. Moreover, the curve implied the appropriate loop count for the model optimization. Moreover, we evaluated the scenario size setting in Appendix D.2.

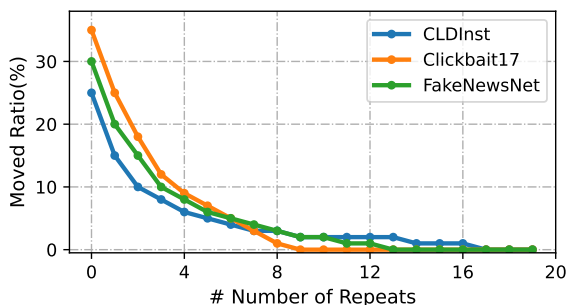


Figure 6: Evaluation on scenario separability based on the ratio of moved samples.

### 3.6 Case Studies

Furthermore, we conducted case study to see the actual effect of our model. Given a dataset like *Clickbait17*, we predicted a score for each test sample. The highest score indicated the scenario it belonged to. After classifying all samples, we selected 5 scenario sets with the largest size of samples. We randomly chose 20 bait and 20 non-bait samples from each scenario, and visualized their features by the dimension reduction tool *t-SNE*. As shown in Fig.(7), subgraph (a) showed the original features cannot differentiate scenarios; (b) presented the causal invariant features learned by our model. These features would be helpful for prediction, but

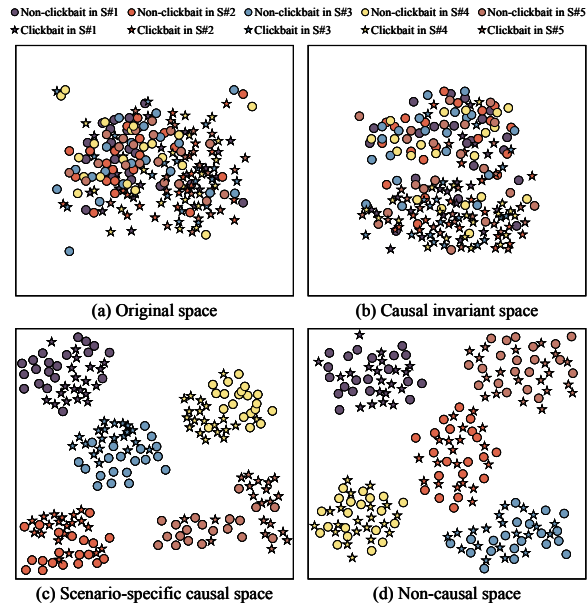


Figure 7: Case study with *t-SNE* visualization on the distribution of samples in several typical scenarios.

they could not distinguish scenarios. The samples formed loosely separated clusters but remained disorganized within each cluster; (c) visualized the scenario-specific causal features. Here, scenarios were separated into distinct clusters well. The features can be roughly divided into two groups, but the inter-class discrimination was insufficient. This demonstrated the need to combine invariant and scenario-specific causal factors for robust prediction; (d) depicted a noise factor. Since the noises in each scenario are unique, they may separate the scenarios but it is weak to identify true/false samples. Overall, these results demonstrated how our model elicited key causal factors and discarded spurious correlations, which enabled reliable prediction.

## 4 Related Work

Clickbait detection is a hot research topic that aims to predict misleading and sensational social posts. The predictive sources can be divided into two categories. The first one is based on social activity. Malicious creators keep posting clickbait and ceaselessly evolve new subspecies to avoid detection. Some researchers propose to examine this bait behavior based on the creators' account profiles (Yang et al., 2012) and activity metadata (Shu et al., 2019b). Nevertheless, some valid creators would produce both high-quality and poor posts, which easily led to false positive predictions. On the other hand, clickbait posts often need to be widely spread on social networks to enhance their



influence. These posts should have remarkable propagation characteristics, such as a unique spread path, a large number of shares but a short reading time due to dissatisfaction, etc. The propagation-based methods had been proposed (Ma et al., 2017), but they cannot tackle new posts due to lack of user feedback (Ma et al., 2018). Thus, such methods often perform high accuracy, but low coverage. Also, they have an alert delay, which would bring a lot of losses and could not support online applications.

Another direction is based on the post’s content. The bait posts often have certain language characteristics, such as delusive words, fallacious phrases, hot topics, raged emotions (Guo et al., 2019), linguistic stylometry (Potthast et al., 2017), etc. To detect them, early works design rules based on various features, such as the semantic (Rony et al., 2017), linguistic (Blom and Hansen, 2015), or multimodal ones (Chen et al., 2015). However, these rules are hand-crafted and non-extensible (Yu et al., 2020). To tackle this issue, current research turns to data-driven neural networks. Various network architectures have been proposed (Krishneth et al., 2023), such as *RNN* (Anand et al., 2017), *CNN* (Agrawal, 2016). More advanced techniques like feature attention (Indurthi et al., 2020) and graph attention mechanisms (Liu et al., 2022a) have been developed. To learn representations from rich unlabeled data (Yu et al., 2023a), some researchers explore the pre-trained language models (*PLMs*), like *BERT* (Devlin et al., 2019) and *RoBERTa* (Liu et al., 2019). They are fine-tuned to fit the classified tasks (Indurthi et al., 2020), so as to use their embedded semantic knowledge (Yu et al., 2023b) to facilitate detection (Yi et al., 2022). Besides, some works point out that the clickbait was not only in the unimodal like the text (Ruchansky et al., 2017) but also in multimodal (Shu et al., 2019a), such as giving an attractive image with an unrelated article. They propose to incorporate more features in multiple modalities (Wang et al., 2018). For model training, there are other studies on domain adaptation (López-Sánchez et al., 2018) and data augmentation (Yang et al., 2019) to solve the data shortage problem by knowledge transfer.

Differently, we found that traditional representations have spurious correlations. Malicious creators would exploit this bug to yield a large number of new subspecies by rewriting the posts’ content, such as adding some valid but unrelated terms or images, and concealing the bait content in a normal post, etc. In this mixed content environment,

the existing model easily misjudges the bait posts due to spurious bias. We thus propose a debiased method to tackle this problem by causal inference.

Causality-inspired methods are a hot research topic in many tasks (Nguyen et al., 2022). To grasp the causality in the tasks, some works proposed the causal loss function (Bagi et al., 2023), while others designed a task-related causal structure (Liu et al., 2022b) to guide decision-making (Lv et al., 2022). They pointed out the value of a good representation for the model’s performance (Bronakowski et al., 2023). Some studies propose to capture the salient representations by feature engineering (Yu et al., 2018) or noise filtering (Wang et al., 2021), but they ignore spurious correlations. In contrast, we propose to elicit key factors to eliminate this bias.

## 5 Conclusion

This paper studied the task of detecting clickbait in multimodal social media posts. Existing methods learned shallow features that introduced spurious correlations, rather than capturing inherent factors that caused clickbait. Malicious creators used this spurious bias to form new bait subspecies by rewriting the posts with tricks, such as disguise with valid content, leading to misjudgment and poor robustness. To tackle this problem, we proposed a new debiased framework by causal representation inference. In detail, we first extracted multimodal features, including textual, visual, linguistic, cross-modal, and creator profile features. By causal inference with structural constraints, we then disentangle them into three latent factors, including the invariant factor that indicated inherent bait intentions, a causal factor for a certain scenario, and noise factor. Based on invariant and causal factors, we can build a robust model. Moreover, we propose a data augmentation technique to reduce training costs. Experimental results on three popular datasets shown the effectiveness of our model.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (62276279, 62372483, 62472455, 62102463, U2001211, U22B2060), Guangdong Basic and Applied Basic Research Foundation (2024B1515020032), Research Foundation of Science and Technology Plan Project of Guangzhou City (2023B01J0001, 2024B01W0004), and Tencent WeChat Rhino-Bird Focused Research Program (WXG-FR-2023-06).

## Limitations

Detecting clickbait posts on social media is a challenging task. In real application, this task comes up against issues such as biases, multimodal content, evolving bait subspecies, few labeled resources, etc. Moreover, malicious creators would constantly disguise the bait posts with tricks such as replacing confusing terms or images pictures to escape detection. This replaced content is full of spurious correlations which would make the model misjudgment. That would seriously affect the model's robustness. We thus propose a debiased model, which can find the posts with inconsistency on the thumbnail-article and headline-article pairs, such as this thumbnail and headline are catchy and sensational, but irrelevant to the article. Although the current model is effective, there is still room for improvement. For example, it does not verify whether the events described in the post are true or fake. Also, it does not cover the detection of video bait posts. Fake content detection is another task and remains an open challenge. A possible solution is to build a trusted knowledge base. We will investigate more data modalities in future work.

## Ethics Statement

The technology proposed in this paper can be used to filter the bait posts. That can improve the user experience and purify the online environment. Unlike traditional methods based on shallow features, our model is more robust by eliminating spurious bias. Excluding the misuse scenarios, there are few or even no ethical issues with this technology. However, it is essentially a classification method, and the classified results may be misused. For example, it may be abused by malicious persons to filter out the posts created from certain commercial competitors unfairly. This problem can be addressed by analyzing the source and publisher of the posts.

## References

- Basant Agarwal, Ajay Agarwal, Priyanka Harjule, and Azizur Rahman. 2023. Understanding the intent behind sharing misinformation on social media. *In Journal of Experimental & Theoretical Artificial Intelligence*, 35(4):573–587.
- Amol Agrawal. 2016. Clickbait detection using deep learning. *In Proceedings of the 2016 2nd international conference on next generation computing technologies (NGCT)*, pages 268–272, Dehradun, India.
- Cleber Alcântara, Viviane Moreira, and Diego Feijo. 2020. Offensive video detection: dataset and baseline results. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4309–4319, Marseille, France.
- Ankesh Anand, Tanmoy Chakraborty, and Noseong Park. 2017. We used neural networks to detect clickbaits: You won't believe what happened next! *In Proceedings of the 39th Advances in Information Retrieval European Conference on IR Research, ECIR 2017*, pages 541–547, Aberdeen, UK.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.
- Shayan Shirahmad Gale Bagi, Zahra Gharaee, Oliver Schulte, and Mark Crowley. 2023. Generative causal representation learning for out-of-distribution motion forecasting. *arXiv preprint arXiv:2302.08635*.
- Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *In Journal of Pragmatics*, 76:87–100.
- Mark Bronakowski, Mahmood Al-khassaweneh, and Ali Al Bataineh. 2023. Automatic detection of clickbait headlines using semantic analysis and machine learning techniques. *In Journal of Applied Sciences*, 13(4):2456.
- Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as "false news". *In Proceedings of the 2015 ACM Workshop on Multimodal Deception Detection*, pages 15–19, Seattle, Washington, USA.
- Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. 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)*, pages 627–638.
- Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: a survey of deep learning techniques. *In Journal of Social Network Analysis and Mining*, 13(1):1–22.
- Claudia Ioana Coste and Darius Bufnea. 2021. Advances in clickbait and fake news detection using new language-independent strategies. *In Journal of Communications Software and Systems*, 17(3):270–280.
- J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the The North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, Minneapolis, MN, USA.

- Varun Dogra, Sahil Verma, NZ Jhanjhi, Uttam Ghosh, Dac-Nhuong Le, et al. 2022. A comparative analysis of machine learning models for banking news extraction by multiclass classification with imbalanced datasets of financial news: Challenges and solutions. *In Journal of International Journal of Interactive Multimedia & Artificial Intelligence*, 7(3).
- Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728*.
- Yu-i Ha, Jeongmin Kim, Donghyeon Won, Meeyoung Cha, and Jungseock Joo. 2018. Characterizing Clickbaits on Instagram. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pages 92–101, Stanford, CA, USA.
- Arafat Hossain, Md Karimuzzaman, Md Moyazzem Hossain, and Azizur Rahman. 2021. Text mining and sentiment analysis of newspaper headlines. *In Journal of Information*, 12(10):414.
- Vijayaradhi Indurthi, Bakhtiyar Syed, Manish Gupta, and Vasudeva Varma. 2020. Predicting clickbait strength in online social media. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4835–4846, Barcelona, Spain (Online).
- A Krishneth, JD Dharaneesh, S Jisnu, and D Sivaganesan. 2023. Web-plugin to detect clickbait in news articles using rnn and lstm. In *Proceedings of the 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 415–420, Coimbatore, India.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, Washington, DC, USA.
- Shu-Hsien Liao, Retno Widowati, and Yu-Chieh Hsieh. 2021. Investigating online social media users' behaviors for social commerce recommendations. *In Journal of Technology in Society*, 66:101655.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, Venice, Italy.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the Computer Vision—ECCV 2014*, pages 740–755, Zurich, Switzerland.
- Tong Liu, Ke Yu, Lu Wang, Xuanyu Zhang, Hao Zhou, and Xiaofei Wu. 2022a. Clickbait detection on wechat: A deep model integrating semantic and syntactic information. *In Journal of Knowledge-Based Systems*, 245:108605.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. 2022b. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17081–17092, New Orleans, LA, USA.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, Montreal, QC, Canada.
- Daniel López-Sánchez, Jorge Revuelta Herrero, Angélica González Arrieta, and Juan M Corchado. 2018. Hybridizing metric learning and case-based reasoning for adaptable clickbait detection. *In Journal of Applied Intelligence*, 48:2967–2982.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *In Journal of Advances in neural information processing systems*, 32.
- Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, New Orleans, LA, USA.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 708–717, Vancouver, Canada.
- Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1989, Melbourne, Australia.
- Peya Mowar, Mini Jain, Ruchika Goel, and Dinesh Kumar Vishwakarma. 2021. Clickbait in youtube prevention, detection and analysis of the bait using ensemble learning. *CoRR*, abs/2112.08611.
- Prithwiraj Mukherjee, Souvik Dutta, and Arnaud De Bruyn. 2022. Did clickbait crack the code on virality? *In Journal of Academy of Marketing Science*, 50(3):482–502.

- Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany.
- Toan Nguyen, Kien Do, Duc Thanh Nguyen, Bao Duong, and Thin Nguyen. 2022. Front-door adjustment via style transfer for out-of-distribution generalisation. *arXiv preprint arXiv:2212.03063*.
- Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th international conference on computational linguistics*, pages 1498–1507.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning*, pages 8748–8763, Virtual Event.
- Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, page 232–239, New York, USA.
- Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, Singapore.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*, pages 395–405, New York, USA.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020a. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. In *Journal of Big data*, 8(3):171–188.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020b. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 626–637.
- Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019b. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439, Vancouver, British Columbia, Canada.
- Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.
- Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021*, pages 1288–1297, Virtual Event, Canada.
- Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 849–857, London, UK.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.
- Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37, pages 2048–2057, Lille, France.
- Kapil Kumar Yadav and Nipun Bansal. 2023. A comparative study on clickbait detection using machine learning based methods. In *Proceedings of the 2023 International Conference on Disruptive Technologies (ICDT)*, pages 661–665, Greater Noida, India.
- Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Pro-*

ceedings of the ACM SIGKDD Workshop on Mining Data Semantics, New York, USA.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *CoRR*, abs/1907.07347.

Xiaoyuan Yi, Jiarui Zhang, Wenhao Li, Xiting Wang, and Xing Xie. 2022. Clickbait detection via contrastive variational modelling of text and label. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 4475–4481, Vienna, Austria.

Jianxing Yu, Xiaojun Quan, Qinliang Su, and Jian Yin. 2020. Generating multi-hop reasoning questions to improve machine reading comprehension. In *Proceedings of the World Wide Web Conference, WWW*, pages 281–291, Taipei, Taiwan.

Jianxing Yu, Qinliang Su, Xiaojun Quan, and Jian Yin. 2023a. Multi-hop reasoning question generation and its application. In *Journal of IEEE Transactions on Knowledge and Data Engineering*, 35(1):725–740.

Jianxing Yu, Shiqi Wang, Libin Zheng, Qinliang Su, Wei Liu, Baoquan Zhao, and Jian Yin. 2023b. Generating deep questions with commonsense reasoning ability from the text by disentangled adversarial inference. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 470–486, Toronto, Canada.

Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 649–658, Lyon, France.

Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. 2020. Fake news early detection: A theory-driven model. In *Journal of Digital Threats: Research and Practice*, 1(2):1–25.

Jie Zhu, Huabin Huang, Banghuai Li, and Leye Wang. 2021. E-crf: Embedded conditional random field for boundary-caused class weights confusion in semantic segmentation.

Yi Zhu, Han Wang, Ye Wang, Yun Li, Yunhao Yuan, and Jipeng Qiang. 2023. Clickbait detection via large language models. *arXiv:2306.09597*.

## A Multimodal Feature Extraction

**Visual Features:** *Swin-T* (Liu et al., 2021) is a transformer-based model that is good at capturing hierarchical features in images based on shifted window self-attention.

**Cross-modal Features:** We employ a *CT* Transformer (Lu et al., 2019) to compute the cross-modal features between the text and image. Given the textual encoding  $\mathbf{T}^b$  and visual one  $\mathbf{V}^s$  with their data type tags, *CT* can well encode their

matching features by a multi-headed attention network. It outputs a visual-aware text feature  $\mathbf{F}^{vt}$  and a text-aware visual feature  $\mathbf{F}^{tv}$ , respectively by  $\mathbf{F}^{vt} = CT((\mathbf{T}^b \mathbf{W}^t), (\mathbf{V}^s \mathbf{W}^v))$ ,  $\mathbf{F}^{tv} = CT((\mathbf{V}^s \mathbf{W}^v), (\mathbf{T}^b \mathbf{W}^t))$ , where  $\mathbf{W}$  are weight matrices. These encodings can better reflect bait behavior from multiple views.

**Linguistic Features:** To capture the semantics and bait patterns of the post, we extract six kinds of features based on post content, i.e., thumbnail, headline, and article body:

(1) *ArticleBody-thumbnail disparity:* Creators may exploit tricks, such as using attractive thumbnails that are irrelevant to the article body, to mislead readers into clicking. To calculate this cross-modal inconsistency, we encode them into a uniform mathematical space by a pre-trained *CLIP* model (Radford et al., 2021), as  $[b^c, t^c] = CLIP(b, t)$ , where  $b$  denotes the body text or figure in the article,  $t$  refers to the thumbnail picture,  $b^c$  and  $t^c$  are their encoded vectors, respectively, where  $d_c$  is the length. Previous works (Zhu et al., 2021) show that *CLIP* has a good ability to characterize the mutual relations of various modalities.

(2) *ArticleBody-headline disparity:* To attract readers, creators often use lure headlines to form a curiosity gap, regardless of their relevance to the articles. To compute their relevance, we first embed each textual input via *BERT* and encode its context by a bidirectional *GRU* with an attention layer. We then feed these two inputs into a *Siamese* net (Neculoiu et al., 2016) that is good at learning similarity, as  $sim_{bh} = Siamese(b, h)$ , where  $b$  denotes the body text,  $h$  refers to the headline.

(3) *Thumbnail-headline disparity:* When the headline doesn't match the thumbnail, it will create a curiosity gap and cause misleading clicks. To understand the thumbnail, we generate its caption based on a neural model (Xu et al., 2015) trained on *MS COCO* (Lin et al., 2014) dataset. We then compare the generated caption and headline by cosine matching of their *BERT* embeddings.

(4) *Sentiment of headline:* The headline often expresses strong feelings to form an emotional curiosity gap. To analyze such feelings, we use a sentiment classifier (Hossain et al., 2021) which outputs both polarity (positive/negative/neutral) and intensity (strength). We normalize their values to the range  $[0, 1]$  and obtain a feature vector.

(5) *Lexical analysis of headline:* To grab the reader's attention, the headline usually uses provoking clues, such as the overuse of numbers, question

marks, exclamation marks, and capital letters to emphasize the shock, or using emojis to indicate the funny emotion. We thus design a count-based vector to quantify the features like capitalized letters, question marks, punctuation marks such as ‘~,!’ , emojis, assertive verbs, factive verbs, hedges, implicative verbs, etc.

(6) *Baitness of headline*: Another striking peculiarity is the lure verbalism, for example, using abbreviations like “OMG” i.e., *oh my god* to express surprise, “LOL” i.e., *laughing out loud* to describe humor, “ROFL” i.e., *rolling on the floor laughing*. The use of celebrity names or pornographic words (Alcântara et al., 2020) like “nudes.” We thus extract features by counting the number of celebrities, slang words such as “OMG, WTF”, porn words like “sexy”, captivating phrases like “you wouldn’t believe,” “shocking.”

## B Experiment Settings

For evaluation, we reimplemented each baseline with default settings. For fair comparisons, we conducted ten runs and showed the average results.

**Ours:** Our model was trained on four 24 GB *Nvidia RTX 3090 GPUs*. Based on *HuggingFace PyTorch* API (Wolf et al., 2020), we encoded the images based on the *Swin<sub>Tiny</sub>* model pre-trained on *ImageNet* with 4 layers and a hidden dimension of 96. The transformer had a window size of  $7 \times 7$  and 6 attention heads. For text encoding, we leveraged the *BERT<sub>base</sub>* model with 6 transformer layers and a hidden size of 768. The classifier was a 3-layer *MLP* with *ReLU* activations and a hidden size of 512. The *Adam* optimizer was used with a learning rate of  $2e - 5$ . We trained for 10 epochs, with the batch size of 64. For contrastive learning, we used a projection head with 256 units and a temperature of 0.1. The mask generator was a 2-layer *MLP*. The learning rate for *m* was obtained by greedy search on [0.01, 0.001, 0.0001]. The trade-off  $\alpha$  and  $\beta$  were set as 2 and 0.1 for *CLDInst*, 1 and 0.1 for *Clickbait17* and 1 and 0.01 for *FakeNewsNet*, respectively. The number of scenarios  $|S|$  was set as 10 for *CLDInst*, 15 for *Clickbait17* and 15 for *FakeNewsNet*, respectively. The iteration *T* was set as 20 initially.

**dEFEND:** We used pre-trained *GloVe* embeddings of dimension 100 to represent words. We incorporated both bidirectional *GRU* layers for sequential data processing and custom attention mechanisms to focus on relevant parts of the text.

The max sentence length and sentence count were set to 120 and 50, respectively. We utilized the *RMSprop* optimizer with a learning rate of 0.001. We trained for 10 epochs, with the batch size of 20.

**HPFN:** We extracted the features from the hierarchical propagation network, and then employed a series of ensemble classifiers to predict the results, *Gaussian Naive Bayes*, *Logistic Regression* with ‘*lbfgs*’ solver, *Decision Tree*, *SVM* with a linear kernel. Also, we used *Random Forest* with 50 estimators for detailed training and 100 estimators for broader model performance analysis. Data normalization was achieved through *StandardScaler*. We used *Extra Trees Classifier* to build a forest with 100 estimators and computed the feature weighting.

**MCAN:** Textual features were extracted using the *BERT* model, and visual features from images were obtained through the *VGG-19* network. The model incorporated a novel fusion approach with multiple co-attention layers to effectively integrate textual and visual features. Specifically, the model was designed with ‘*s-fc*’, ‘*f-fc*’, and ‘*t-fc*’ layers. Each had a hidden size of 256, and a ‘*p-fc*’ layer with a hidden size of 35. The dimensions *d*, *m*, and *dff* were set to 256, 4, and 512, respectively. Training was conducted for 100 epochs with early stopping, utilizing *Adam* optimizers. The *VGG-19* and *BERT* parameters were frozen to prevent overfitting. Parameters were optimized by grid searching, with accuracy as the selection criterion.

**CPDM:** We embedded the input text by the *BERT* model. The dimension of the last hidden layer was set to 768. For the visual input, we embedded it by the *ResNet50* pre-trained model with the output layer dimension of 2048. We then fused these multimodal features and trained an ensemble classifier, i.e., *Random Forest*. It consisted of six base classifiers, i.e., *K-Nearest Neighbours*, *Gaussian Naive Bayes*, *Multi-Layer Perceptron*, *Support Vector Machines*, *Extreme Gradient Boosting*, and *Logistic regression*. We used *Adam* as the optimizer with a learning rate of  $10^{-4}$  and a batch size of 128. The training epoch was set to 120.

**CCD:** We trained the model for 200 epochs. The initial learning rate for the *Adam* optimizer was tuned in [1e-5, 1e-3]. For the confounder dictionary  $\mathbf{D}_u \in \mathbb{R}^{N \times d_u}$ , *N* is 18 (*Anger*, *Anxiety*, *Assent*, *Causation*, *Certainty*, *Differentiation*, *Discrepancy*, *Feel*, *Hear*, *Insight*, *Negative emotion*, *Netspeak*, *Nonfluencies*, *Positive emotion*, *Sadness*, *See*, *Swear words*, *Tentative*), and *d<sub>u</sub>* was set to 4. For the scaled dot-product attention, the scaling

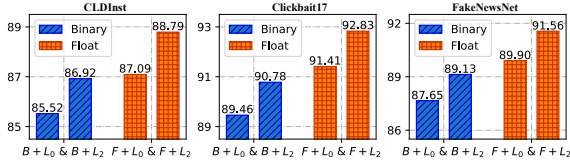


Figure 8: Evaluation the impact of  $\mathbf{m}$  settings on ACC.

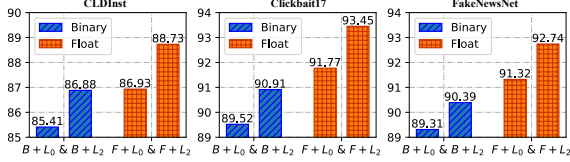


Figure 9: Evaluation the impact of  $\mathbf{m}$  settings on PRE.

factor  $d_m$  was set to 256. We followed the original settings to tune the trade-off hyperparameters  $\alpha$  and  $\beta$  by grid search in  $\{0, 0.1, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5\}$ . And we finally set  $\alpha = 3$  and  $\beta = 0.1$ .

**VLP:** Due to the storage limitation, we used the first half of the *YT-Temporal* 180M. We trained the model using the *AdamW* optimizer with a base learning rate of  $1e-4$  and weight decay of  $1e-2$ . The learning rate was warmed up for 10% of the training steps and was decayed linearly to zero for the rest of the training. For pre-training, we trained *All-in-one-S* and *All-in-one-B* for 200K steps with a batch size of 8 per GPU. For *All-in-one-Ti*, we pre-trained for 100K steps with a batch size of 16 per GPU. We adopted mixed precision technique to speed up the training process. As the domain gap between pre-train dataset and downstream visual dataset is large, we used batch size of 512 and trained with 100 epochs.

## C Training Procedure

Our overall training process is summarized in Algorithm 1.

## D Additional Evaluations

Due to the page limit, we showed additional experiments as follows, including the evaluations of the mask and scenario mechanism, case analysis, and data augmentation technique.

### D.1 Additional Study of the Mask Mechanism

The results in terms of accuracy, precision, and recall metrics were presented in Fig.(8), Fig.(9), and Fig.(10), respectively. All the results validated our design of using a soft float mask regularized by  $L_2$  norm, which could perform best. The mask en-

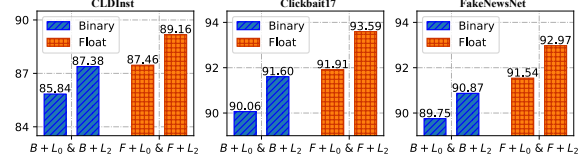


Figure 10: Evaluation the impact of  $\mathbf{m}$  settings on REC.

---

### Algorithm 1: Our model’s training process.

---

**Data:** Training data  $\mathcal{D}^{tr}$

**Result:** Optimal detector  $\hat{\mathcal{F}}$

- 1 Random assign scenario for each  $x_i \in \mathcal{D}^{tr}$ ;
  - 2 //  $T$  rounds of alternating optimization;
  - 3 **for**  $t \leftarrow 1$  **to**  $T$  **do**
  - 4     **while not converged do**
  - 5         Optimize  $\Phi_s$  via Eq.(2) for scenario division;
  - 6         Compute  $s(i)$  via Eq.(3) for samples reallocation;
  - 7     **end**
  - 8     Learn *IRM* loss  $\mathbf{m}$  via Eq.(1) to elicit an invariant factor;
  - 9     Optimize the contrastive loss  $\xi$  via Eq.(4) to separate scenario-specific causal factor and non-causal one;
  - 10 **end**
  - 11 Optimize Eq.(5) to learn a clickbait detector  $\hat{\mathcal{F}}$  based on the elicited invariant and causal factors, avoiding spurious bias
- 

abled automatic differentiation of feature relevance for clickbait detection.

To further analyze the mask  $\mathbf{m}$ , we visualized its per-dimension values in Fig.(11). For each dataset, we randomly sampled 100 cases and then analyzed their learned mask weights. This weight captured the importance of each kinds of feature. The value of this weight was between  $[0,1]$ . If we plotted the histogram according to the weights of these 100 samples, where the x ordinate is the value of the weight and the y ordinate is its corresponding number of samples. For example, if there was 17 samples had the weight 0.4, then x ordinate was 0.4, the y ordinate was 17. If the weight values of most samples were high, this kind of feature was considered to be more important. From the results, we observed that there were more samples with high weight for the kinds of cross-modal, linguistic, and profile features. This reflected their ability to capture bait behavior more effectively than the single modal features such as text and visual ones.

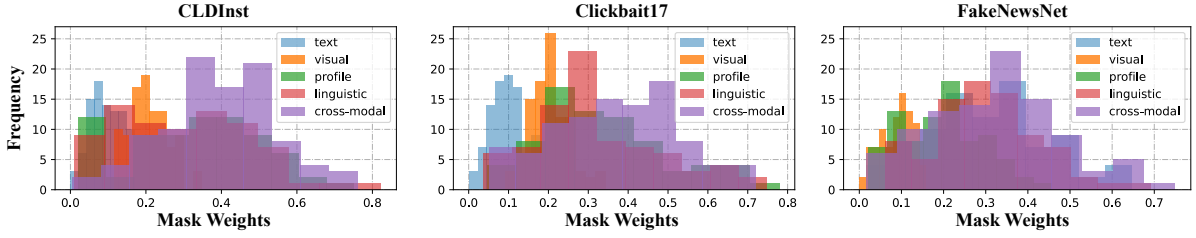


Figure 11: Visualization of the mask  $m$  on three datasets. The importance distribution of each kind of feature.

## D.2 Study of the Scenario Size Setting

We evaluated the impact of scenario size setting on the overall performance. Empirically, the small size was not enough to distinguish complex clickbait patterns, while a large one may split some relevant samples into redundant groups. To explore the best settings, we tuned the scenario size from 1 to 25, with 1 as an interval. Considering the computational resources, we tested the size with no larger than 25. The performance change curves in terms of accuracy, F1 score, precision, and recall were displayed in Fig.(12). We observed that with the increment of size, the performance had an ascending trend, after reaching the peak, it turned to decline. The optimal sizes for the three datasets were 10, 15, and 15, respectively.

## D.3 Qualitative Analysis of Test Cases

Moreover, we analyzed the test cases to infer the actual running effect of our model. As exhibited in Table 3, we made the correct prediction for the first post. This article used a cartoon-style thumbnail. Our model could judge the authenticity of this article through the learned textual features and visual consistency, without being misled by the attractiveness of the title or image itself. In the second example, the headline used exaggerated words like “*Bid Rupert Murdoch farewell by reliving his most controversial tweets.*” It created a curiosity gap which people expected to reveal the tweets immediately. But the image of the post only contained a side-sitting photo of “*Rupert Murdoch*”, which had nothing to do with the headline. The body text actually simply reviewed some tweets posted by *Rupert Murdoch* on Twitter, without delivering the implied sensational effect as hinted by the headline. For this clickbait post with inconsistency and irrelevance, other baselines might rely too much on the surface word matching between the headline and body text, without fully understanding their deeper semantics, leading to the wrong judgment. In contrast, our model can work well. These results showed the ex-

tracted invariant and scenario-specific factors were effective in distinguishing bait articles.

## D.4 Study of Data Augmentation Technique

In real applications, training data is usually insufficient or even not. To address this problem, we develop a data augmentation technique which collected pseudo-labeled samples from the historical data in social media, e.g. *WeChat*. We found that although the social behavior-based approach needed extra cold start time, its accuracy was high if sufficient feedback data was accumulated over time. We thus propose to collect the user behavior of these posts from historical data over a period of time, so as to build a batch of pseudo-labeled data. In detail, we first collected hot posts with more than  $\mu = 100,000$  forward actions per hour. We then designed two heuristic rules to identify clickbait, including those that take less than  $\nu = 10$  seconds viewing duration; and the number of user likes is 0. The remaining posts were viewed as non-bait. The statistics on dataset are shown in Table 4. We used the *Google Translate API* to translate these Chinese posts into English, and got a data-augmented dataset *WeChatCB*. To analyze the quality of this dataset, we randomly selected 500 samples for human evaluation. The result showed that approximately 28% of the posts were disguised type. It indicated that *WeChat*, like other popular social platforms, existed serious deceptive issue.

To verify the effectiveness of this augmentation technique, we simulated a few-shot and zero-shot environment and trained our model on the augmented data. If the quality of this data is equivalent to the human-tagged one, their performance should be comparable. In detail, we retained 10% of the training data from the original dataset and randomly selected data with a size of 90%  $\sim$  190% from the augmented dataset as an additional training set. The selected data included both positive and negative samples. The results on the test sets of *CLDInst*, *Clickbait17*, and *FakeNewsNet*, respec-





Thumbnails	Headlines	Body Texts	True Label	Pred Label
	Anderson Cooper hits Trump with a hard truth about how he'll go down in history	“Seven days after choosing to be the first president ever to incite insurrection against the country and the Constitution he took an oath to defend, the president now has another first to his name,” Cooper said. “The first and only president to ever be impeached twice.”	0	dDEFEND : 1 HPFN : 1 MCAN : 1 CPDM : 1 CCD : 0 VLP : 0 Ours : 0
	Bid Rupert Murdoch farewell by reliving his most controversial tweets	Rupert Murdoch, one of the top media executives in the world, is stepping down from his perch as CEO of 21st Century Fox. . . “Independence Day. Immigration is our history and immigration MUST be our future. Multi-ethnicities and equality under law for all. “Making rich poorer won’t do much. Giving opportunity too poor to become rich is the only way forward.”	1	dDEFEND : 0 HPFN : 0 MCAN : 0 CPDM : 0 CCD : 1 VLP : 0 Ours : 1

Table 3: Qualitative analysis of test cases for various methods. 0 indicates non-clickbait, while 1 denotes clickbait.

Dataset	#Samples	#Clickbait	#non-Clickbait
WeChatCB	70,785	32,418	38,367

Table 4: The statistics on the pseudo-labeled samples.

Datasets	CLDInst		Clickbait17		FakeNewsNet	
	PRE ↓	REC ↓	PRE ↓	REC ↓	PRE ↓	REC ↓
w/o MF	-4.59 ± 0.14	-4.85 ± 0.52	-5.81 ± 0.37	-6.11 ± 0.29	-5.42 ± 0.22	-5.79 ± 0.04
w/o EICF	<b>-5.89</b> ± 0.45	<b>-6.16</b> ± 0.17	<b>-7.79</b> ± 0.11	<b>-8.32</b> ± 0.30	<b>-6.90</b> ± 0.26	<b>-7.23</b> ± 0.19
w/o ESCF	-3.53 ± 0.43	-3.87 ± 0.55	-4.02 ± 0.29	-4.52 ± 0.06	-3.81 ± 0.14	-3.93 ± 0.11
w/o ENF	-3.78 ± 0.35	-4.06 ± 0.12	-4.62 ± 0.51	-4.85 ± 0.43	-4.14 ± 0.06	-4.49 ± 0.40

Table 5: Ablation studies. T-test, p-value < 0.005.

tively are shown in Table 6. Moreover, we also showed the results of using only the augmented samples (from a size of 100% to 200%), but with no original training data. Based on 10% human-tagged data and 150% machine-tagged data, our model obtained better performance than the one based on 100% original training data (i.e., human-tagged data). When in the zero-shot situation, our model still can obtain good performance based on 180% machine-tagged data. The outperformance would be larger when feeding more tagged data. We can infer that the quality of machine tagging is acceptable and satisfactory. Our data augmentation technique is useful to alleviate for the shortage problem of human-tagged resources.

## D.5 Ablation Studies on Other Metrics

As demonstrated in Table 5, we plotted the results in terms of *PRE* and *REC* metrics, respectively.

Data Settings	CLDInst				Clickbait17				FakeNewsNet			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
<b>H 100%</b>	88.79±0.11	88.73±0.04	89.16±0.32	88.94±0.14	92.83±0.27	93.45±0.17	93.59±0.15	93.52±0.20	91.56±0.07	92.74±0.38	92.97±0.13	92.85±0.08
<b>H 10% + M 90%</b>	83.25±0.31	85.99±0.22	84.40±0.05	85.19±0.17	87.95±0.58	90.16±0.03	90.56±0.71	90.36±0.24	86.19±0.36	89.64±0.49	88.99±0.10	89.31±0.12
<b>H 10% + M 110%</b>	85.53±0.06	87.37±0.42	86.66±0.33	87.01±0.50	90.27±0.21	91.84±0.16	91.70±0.47	91.77±0.13	88.06±0.11	91.33±0.27	90.85±0.24	91.09±0.09
<b>H 10% + M 130%</b>	88.22±0.32	88.51±0.24	88.37±0.08	88.44±0.17	92.62±0.14	93.06±0.49	92.86±0.36	92.96±0.10	90.21±0.21	92.28±0.40	92.53±0.29	92.41±0.16
<b>H 10% + M 150%</b>	89.87±0.03	89.82±0.13	89.38±0.51	89.60±0.44	93.80±0.28	93.33±0.24	93.72±0.30	93.52±0.15	92.02±0.44	92.84±0.38	93.05±0.07	92.95±0.32
<b>H 10% + M 170%</b>	91.85±0.24	91.28±0.35	90.51±0.56	90.89±0.17	94.07±0.13	93.71±0.25	94.25±0.08	93.98±0.04	93.00±0.52	93.47±0.55	93.98±0.09	93.73±0.31
<b>H 10% + M 190%</b>	92.15±0.51	92.51±0.22	92.57±0.38	92.54±0.05	94.43±0.34	94.35±0.27	94.64±0.18	94.49±0.44	94.07±0.08	94.37±0.20	94.48±0.54	94.42±0.42
<b>M 100%</b>	82.90±0.56	85.29±0.38	85.42±0.21	85.35±0.60	87.20±0.44	90.05±0.05	88.71±0.19	89.37±0.24	87.89±0.43	89.52±0.51	86.25±0.11	87.85±0.39
<b>M 120%</b>	84.46±0.18	86.46±0.05	86.64±0.14	86.55±0.62	88.30±0.30	90.91±0.11	89.66±0.25	90.28±0.32	89.33±0.35	90.93±0.50	88.03±0.12	89.46±0.29
<b>M 140%</b>	86.22±0.24	87.41±0.30	87.58±0.58	87.49±0.16	90.26±0.29	91.89±0.08	90.62±0.03	91.26±0.29	90.82±0.51	91.51±0.40	89.15±0.05	90.32±0.14
<b>M 160%</b>	87.76±0.62	87.92±0.40	88.30±0.18	88.11±0.09	91.78±0.23	92.75±0.41	91.77±0.57	92.26±0.54	91.98±0.34	92.02±0.50	91.33±0.21	91.67±0.26
<b>M 180%</b>	89.14±0.43	89.20±0.16	88.85±0.22	89.03±0.24	92.54±0.17	93.36±0.62	93.91±0.07	93.63±0.35	92.46±0.26	92.47±0.06	92.30±0.45	92.38±0.10
<b>M 200%</b>	90.11±0.14	90.89±0.39	90.75±0.06	90.82±0.42	94.02±0.64	93.78±0.19	93.58±0.20	93.68±0.18	93.31±0.15	93.08±0.53	93.88±0.33	93.48±0.07

Table 6: Evaluation of the quality of the augmented data. H and M represent the human-tagged data and machine-augmented data which are used to train our method, respectively.

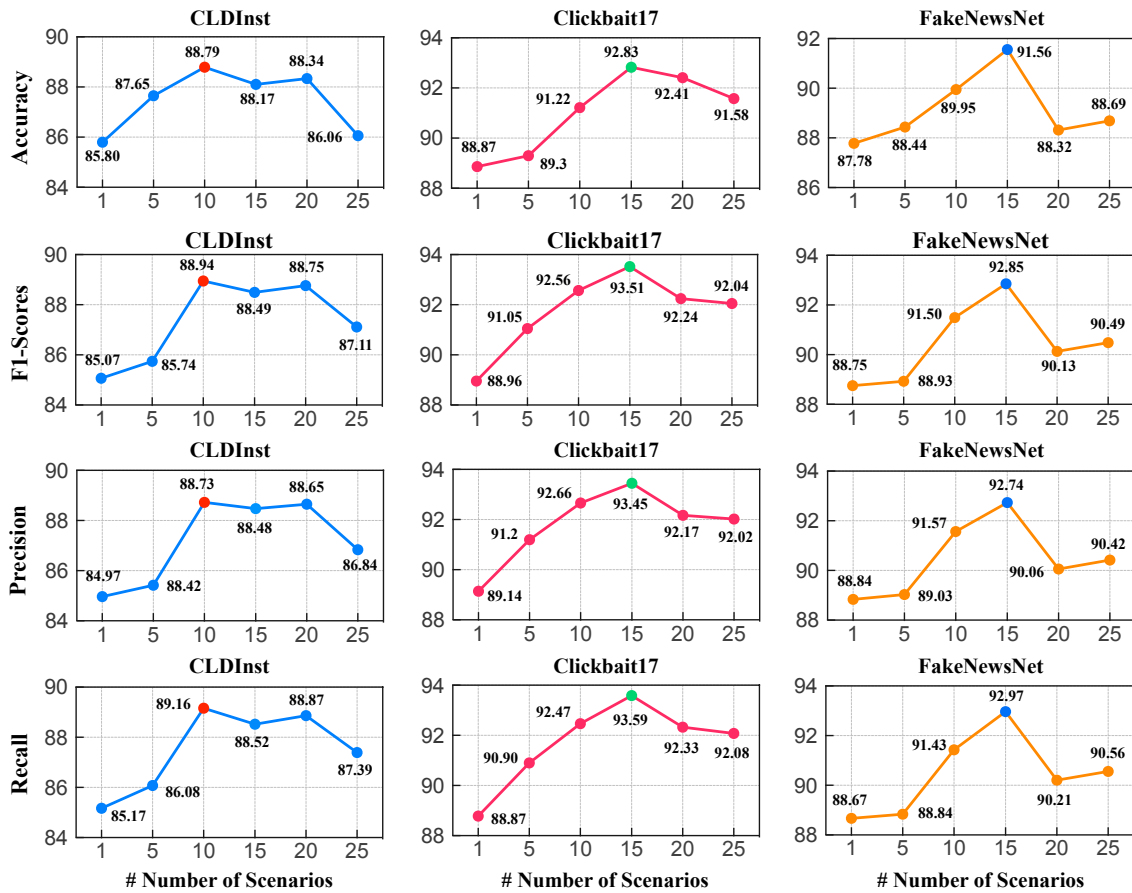


Figure 12: Evaluation of scenario size settings in terms of accuracy, F1 score, precision and recall, respectively.