Revealing Subtle Marketing: A Benchmark and Framework for Social Media Covert Advertisements Detection

Anonymous ACL submission

Abstract

Currently, posting covert advertisements on social media is an increasingly common marketing strategy. This practice will mislead users, which may influence their decisions and cause unfair competition, highlighting the urgent need for effective detection methods. However, research on this topic remains limited. In this study, we formalize the covert advertisement detection task and present the first social media covert advertisement benchmark. The benchmark includes Chinese and English posts collected from two representative social media platforms (Rednote and Instagram) with manually annotated labels. We evaluate several multimodal methods and find that, as covert advertisements can appear within a single modality or through cross-modal interplay, these methods struggle with effective detection and fail to adequately balance single-modal and fused features. To address this challenge, we propose SCAN (**S**ocial-media **C**overt **A**dvertisement Detection using Multi-view **N**etwork), a framework that leverages cooperative training to better balance and utilize both single-modal and fused features. Our results show that SCAN can further advance covert advertisement detection performance. We believe our benchmark and method will contribute to future research in social media covert advertisement detection.

1 Introduction

Advertisements on social media can effectively help businesses expand their market and attract consumers (Keller et al., 2010). However, as the number of advertisements increased, people have become accustomed to and skeptical of them, reducing their effectiveness (Petty and Andrews, 2008). To address this issue, covert advertisements have been developed and widely adopted (Wojdynski and Evans, 2020).

Covert advertisements disguise marketing content as seemingly normal user-generated content,

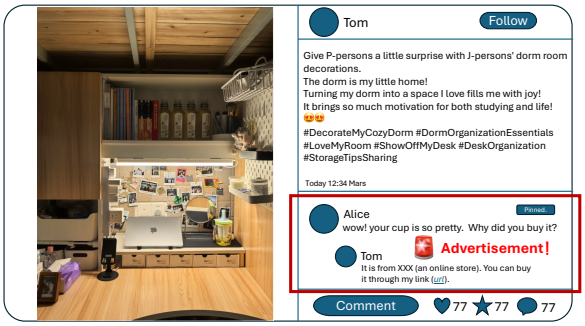

Figure 1: A typical example of covert advertisement. In this example, neither the image nor the body of the text is shown to contain a sales pitch, but in the pinned comment, the author shows a strong intent to lead.

without any indication to alert users (Pierre, 2023), while traditional advertisements usually have clear marks to indicate their marketing intentions, such as being endorsed by a public figure or with explicit advertising slogan (Shamdasani et al., 2001).

Covert advertisements often disguise themselves as common posts with hidden branding, promotional remarks, or boosted visibility from user actions like comments and shares (Amazeen, 2023). As shown in Figure 1, covert advertisements are designed to closely resemble ordinary posts, making them difficult to detect. If consumers do not realize the persuasive intent behind such content, they are more likely to trust the content, leading to unfair marketing practices (Rozendaal et al., 2010). This misleads consumers and gives some advertisers an unfair edge, which will hurt fair competition. This also makes content regulation on platforms more difficult (Austin and Newman, 2015).

Despite the significant negative impact of covert advertisements, research on this topic remains limited.

Since covert advertisements contain both images and text, the multimodal model could be a powerful way to detect covert advertisements. Detecting

covert advertisements in social media remains a tough task, even with multimodal models. First, there is no clear definition or guidance for the task of identifying "covert advertisements", which complicates research and method development. Second, the lack of benchmarks prevents the evaluation and comparison of different approaches. Finally, covert advertisements are often highly subtle and can appear in various places, such as images, text, or comments, or even through the interplay of different modalities. This complexity makes detection difficult, as relying on a single modality may be insufficient, and integrating features across modalities may risk the loss of key single-modality details that are often critical for identification. These challenges make the task complex and demanding.

Our contributions to tackle these challenges are:

- **Task:** We are the first to introduce the task of detecting covert advertisements across two social media platforms and provide assessment guidelines for defining such advertisements. This task aims to improve social media regulation and offer a clear path for future research.

- **Benchmark:** We introduce a benchmark composed of covert advertisement posts from the Chinese social media platform (Rednote[1]) and the English social media platform (Instagram[2]). This dataset is manually annotated and includes challenging samples, such as some special character or symbol[3] in RedNote used to evade detection. We also evaluate the performance of state-of-the-art general-purpose multimodal models and domain-specific methods on our benchmark.

- **Approach:** We develop SCAN, a multi-view social media covert advertisement detection network. It uses cooperative training to integrate single-channel and multi-channel views to ensure the balance between single-modal and fused features. Experiments show that the SCAN has a better performance than existing methods.

[1]RedNote (https://www.xiaohongshu.com) is one of the most popular social platforms in China, with over 120 million daily active users and more than 300 million monthly active users.

[2]Instagram (https://www.instagram.com) has more than 500 million daily active users and 2.11 billion monthly active users (https://www.demandsage.com/instagram-statistics/).

[3]We show some examples in Appendix B.2

2 Related Work

Social Media Content Moderation. Moderating social media content is crucial for ensuring fair business practices, maintaining social order, and safeguarding mental health (Gongane et al., 2022). Current research focuses on identifying various types of harmful content, including hate speech (Ayo et al., 2020), fake news (Sheng et al., 2022), rumors (Ahmed et al., 2017), cyberbullying (Gillespie, 2020), toxic content, and child abuse material (Nahmias and Perel, 2021). Hate speech detection often combines text analysis with social network analysis (Nagar et al., 2023), while fake news detection involves verifying the authenticity of news by comparing similar content (Sheng et al., 2022). Rumor and cyberbullying detection, on the other hand, predominantly leverage NLP methods to analyze textual data (Bharti et al., 2022; Yan et al., 2024). Existing research highlights diverse strategies to tackle these significantly harmful contents, but regulatory efforts have largely overlooked issues like covert advertisements, which, while less direct in their impact, lead to unfair business practices and improper competition. Addressing such challenges is essential for fostering a fair and trustworthy social media environment.

Social Media Dataset. Existing datasets in related domains can be broadly divided into two categories. The first category focuses on traditional advertisements, such as (Li et al., 2022), which collected 20K official Facebook ads to predict revenue, and (Hussain et al., 2017), which compiled 64K advertisement images and 3K videos. Similarly, (Liang et al., 2021) gathered 1K advertisement images to analyze user visual attention, and (Liu et al., 2020) collected 48K textual Chinese advertisement posts to assess legality. The second category includes datasets of user-generated social media content unrelated to advertising. For instance, (Turcan and McKeown, 2019) curated 190K Reddit posts for stress detection, and (Guo et al., 2023) annotated 5K Twitter posts for sentiment analysis. Others, like (Fung and Ji, 2022) and (Santia and Williams, 2018), collected multimodal posts from Weibo and Facebook, focusing on geopolitical events and news authenticity, respectively. While these datasets cover diverse topics, none address the detection or analysis of covert advertisements in social media. To the best of our knowledge, we are the first to propose a benchmark dataset specifically for this task.

Multi-modal Detection. The transformer architecture (Vaswani, 2017), initially developed for NLP tasks, revolutionized the field by introducing attention mechanisms. Pre-trained models like BERT (Devlin, 2018) and GPT-3 (Brown, 2020) showcased the power of large-scale pre-training, with GPT-3 excelling in diverse tasks without fine-tuning. Recent advancements in multi-modal models, such as Qwen-2-VL (Wang et al., 2024), Llava-1.5 (Liu et al., 2023), DeepSeek (Lu et al., 2024), Mini-GPT4 (Zhu et al., 2023), Glm-4v (GLM et al., 2024), and Chat-UniVi (Jin et al., 2024), further demonstrate the effectiveness of combining visual and textual data, highlighting their potential for tasks requiring cross-modal comprehension. Building on this foundation, hierarchical multimodal fusion techniques like those employed in MSD (Cai et al., 2019) have been explored, integrating text, image, and image attribute features for tasks like fake news detection. To enhance cross-modal interactions, methods such as HMCAN (Qian et al., 2021) split textual information into multiple dimensions and utilize attention mechanisms to model its interplay with visual features. Advanced frameworks like CLIP-GCN (Zhou et al., 2024) leverage CLIP for cross-modal feature extraction, adversarial networks for domain adaptation, and graph neural networks to address emergent fake news scenarios. Furthermore, hybrid approaches like LDSF (Ding et al., 2022) have introduced late fusion models that integrate outputs from text, image, and audio classifiers through an exponential fusion function, demonstrating their utility in spam email filtering and similar tasks. We have evaluated these methods in Section 5.

3 Task and Assessment Guidance

Covert advertisement has been around for decades, almost as long as modern advertising practices (Cameron et al., 1996; Erjavec, 2004). However, there is no universally recognized definition of it. To address this, we summarize some of the broad understandings from existing research and present our definition, which will be used throughout this study.

In a narrow definition, covert advertisement refers to the creation of promotional content designed to resemble regular content, deliberately concealing its advertising nature and thereby potentially misleading consumers (Nelson et al., 2009; Kim et al., 2019; Wojdynski and Evans, 2016). This definition highlights two key features: covertness and intent. First, covert advertisement blends seamlessly with the type of content that consumers expect to engage with, making it difficult to identify as promotional. Second, covert advertisement is intentionally designed to subtly influence consumer behavior without being overtly acknowledged This perspective has been adopted in various studies on advertising effectiveness (Wojdynski and Evans, 2020; Nelson et al., 2009; Kim et al., 2019).

A broader definition of covert advertisement may include content that has more than a purely commercial intent, such as political satire (Boerman et al., 2017) or opinionated persuasive content (Campbell and Evans, 2018). In this paper, following (Wojdynski and Evans, 2020), we use the narrower definition. Our formal definition of the covert advertisement is as follows:

Definition 1 *Covert advertisement is content deliberately designed to resemble conventional, non-promotional content with the primary aim of subtly influencing the audience's consumption decisions without explicit disclosure of its advertising nature.*

According to our definition, we provide assessment guidance to users. To be short, users should check all parts of the post, such as images, text, and comments, to observe if there is any specific promotional information. We have also summarized several common types of covert advertisements. The full version is shown in Appendix A. We also list some cases that are not recognized as covert advertisements: 1. Content that is clearly non-factual and unlikely to be mistaken for factual information, such as fictional material. 2. Content with clear context that is oppositional and satirical. 3. Entertainment-oriented content whose main motivation is amusement rather than subtle promotion, even if it includes branded elements.

4 Social Media Covert Advertisements Dataset

Our dataset comprises real user posts from RedNote and Instagram, representing Chinese and English social media, respectively. Some statistics are shown in Table 1 and Figure 2.

4.1 Posts Collection

Rednote. Rednote has become a hugely popular platform in China, widely used for sharing information and connecting with others (Tan, 2024). Users

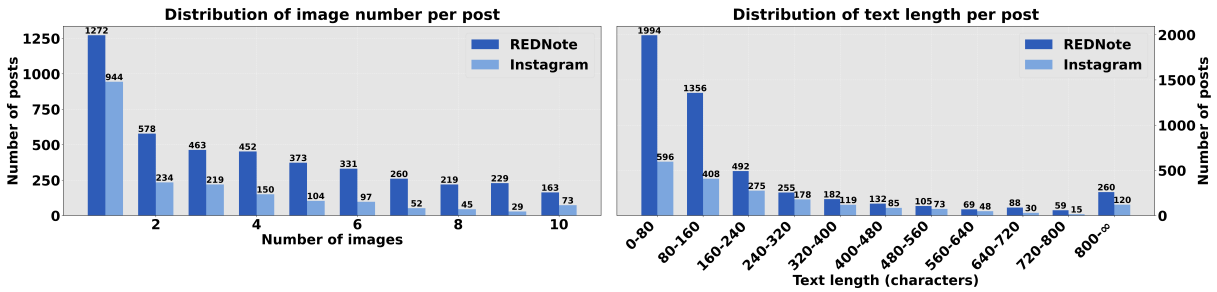

Figure 2: Statistical distributions of our dataset. The left subplot presents the distribution of image count per post), while the right subplot shows the distribution of text length per post.

are drawn to it for reasons like self-expression, entertainment, as well as business promotion. These interactions, influenced by social impact and personal connections, create a lively and engaging community (Kokko, 2023).

To construct an unbiased dataset that accurately reflects real user browsing behavior, we employed three people with brand-new accounts with no browsing history or prior activity. They browsed the homepage daily, clicked on each post, and stayed on it for over five seconds. We collected detailed post information, including images, titles, descriptions, publication dates, and comments, from the browsing history of each user. Notably, our data collection process ignores posts marked with the platform's 'sponsored' tag, which indicates content placed through official channels with a visible promotional logo. We describe the data collection process and further data analysis in more detail in the Appendix B.1.

Instagram. Instagram stands out as the leading platform for influencer marketing (Kim et al., 2020), and several valuable datasets have already been developed to explore this domain (Kim et al., 2020, 2021). Taking into account factors such as dataset size, labeling costs, and accessibility, we select an open-source dataset available on Kaggle[4] as the base dataset for our research. Compared to the Rednote dataset we collected, this base dataset lacks specific comment content but provides other useful information, such as the number of comments, number of likes, and author profile details. We show further data analysis in the Appendix B.1.

4.2 Posts Annotation

To accurately label covert advertisements, we hired three Ph.D. students who were well-versed in using both Redote and Instagram. All of them had

[4] https://www.kaggle.com/datasets/thecoderenroute/instagram-posts-dataset

Table 1: Statistics of our datasets. The sample column shows the number of positive and negative samples in both datasets. The image indicates the average number of images per post. The text column represents the average character length per post, and the comment column represents the average character length per comment.

Dataset	#Sample		#Image	Text	Comment
	Pos.	Neg.			
REDNote	1,091	3,901	5.28	196.63	25.01
Instagram	965	982	2.82	256.35	-

years of experience with these platforms. Their task was to identify whether a post subtly promoted a product, service, or brand without clearly revealing its promotional intent. Given the subtle and implicit nature of covert advertisements, the annotators focused on nuanced cues, such as indirect endorsements, intentional product placements, and persuasive language within seemingly user-generated content. We first trained them according to the assessment guidelines that we mentioned in Section 3. Each post was evaluated independently by at least two annotators. In cases of disagreement, a senior annotator reviewed the post and made the final decision. This multi-step process was designed to enhance annotation consistency and reduce potential biases. To clarify the annotation rules and reduce disputes during the annotation process, posts that fall into the following three categories were directly discarded: (1) Posts where the author explicitly stated that the post had a marketing or promotional intent. (2) Posts with the intent that was too ambiguous to be determined, even by experienced annotators. (3) Posts with incomplete context or content that made assessing intent impossible. In the end, we obtained 4,992 manually annotated posts from Rednote and 1,947 posts from Instagram, which constitute our social media covert advertisements dataset. More details

Table 2: The result of benchmarking existing baselines. ZS refers to zero-shot, where no examples are provided to the model. FS refers to few-shot in-context learning, where the model is given two positive and two negative examples as prompts. FT refers to fine-tuning, where the model is trained on the entire training set and, during inference, does not require additional examples.

Method		Rednote				Instagram			
		ACC ↑	PRE ↑	REC ↑	F1 ↑	ACC ↑	PRE ↑	REC ↑	F1 ↑
Qwen2-VL-7B-Instruct	ZS	0.632	0.731	0.408	0.520	0.665	0.781	0.432	0.556
	FS	0.708	0.712	0.688	0.700	0.652	0.657	0.595	0.624
	FT	0.798	0.873	0.697	0.776	0.836	0.794	0.909	0.848
Llava-1.5-13b-hf	ZS	0.642	0.878	0.330	0.480	0.583	0.814	0.184	0.304
	FS	0.583	0.909	0.184	0.305	0.611	0.797	0.268	0.402
	FT	0.778	0.877	0.651	0.747	0.796	0.828	0.746	0.785
DeepSeek-VL-7B-chat	ZS	0.587	0.583	0.572	0.578	0.665	0.786	0.426	0.553
	FS	0.632	0.630	0.618	0.624	0.670	0.802	0.426	0.557
	FT	0.782	0.798	0.755	0.776	0.804	0.807	0.800	0.804
GLM-4v-plus	ZS	0.627	0.712	0.427	0.534	0.619	0.681	0.446	0.539
	FS	0.600	0.573	0.782	0.662	0.705	0.696	0.727	0.711
MSD		0.743	0.563	0.616	0.589	0.760	0.750	0.758	0.754
HMCAN		0.827	0.789	0.783	0.786	0.802	0.811	0.802	0.800
CLIP-GCN		0.796	0.810	0.773	0.791	0.810	0.815	0.800	0.807
LDSF		0.800	0.780	0.836	0.807	0.790	0.776	0.818	0.797

are shown in Appendix B.2.

5 Benchmarking Existing Baselines

We categorize the existing baselines into two types. The first type comprises popular multimodal large models. The second type consists of classical and effective methods widely used in related fields, such as fake news detection and Spam E-mail detection.

Datasets. We use the dataset introduced in Section 4, splitting it into 70% for training, 10% for validation, and 20% for testing. As shown in Table 1, the dataset is imbalanced, with, for instance, 220 positive samples and 780 negative samples in the Rednote test set. To address this imbalance, we randomly selected negative samples to match the number of positive samples, creating balanced training, validation, and test sets for all experiments.

Metrics. We employed four commonly used metrics for classification tasks: Accuracy (ACC), which measures the overall correctness of the model; Precision (PRE), indicating the proportion of true positives among all positive predictions; Recall (REC), reflecting the model's ability to identify actual positive instances; and F1-score (F1), the harmonic mean of precision and recall.

5.1 Multimodal Large Models (MLMs)

We employ Qwen-2-VL (Wang et al., 2024), Llava-1.5 (Liu et al., 2023), DeepSeek (Lu et al., 2024), and GLM-4v (GLM et al., 2024) for evaluation. These models are evaluated for covert advertisement detection using zero-shot, few-shot, and fine-tuning. In the zero-shot setting, we input the image and text directly for covert advertisement detection without providing examples. In the few-shot setting, we provide two positive and two negative examples as prompts. For the prompt we used, please refer to the Appendix C.1. Note that we also abate the number of samples in this setting and the results are shown in Appendix C.1. For fine-tuning, we use the entire training set and apply LoRA (Hu et al., 2021) with a learning rate of 1e-5 and a batch size of 16 for 5 epochs. We also show the result of different fine-tuning epochs in Appendix C.1. Note that GLM-4v is a closed-source model, so fine-tuning was not performed on it. The results are presented in the upper part of Table 2.

The results demonstrate that current multimodal large models face significant challenges in covert

advertisement detection. For instance, even under the fine-tuning setting, the accuracy of Qwen2-VL and Llava-1.5 on Rednote barely exceeds 79%, while their performance in zero-shot and few-shot settings often lags behind, with accuracy dropping to around 63%. Similarly, on the Instagram dataset, while fine-tuned models like DeepSeek-VL and Qwen2-VL achieve up to 80% accuracy, their performance in zero-shot and few-shot settings remains modest, often approaching random guessing. These findings highlight the inherent difficulty of this task and the limitations of existing models in addressing it effectively.

Furthermore, we notice that the improvement from zero-shot to few-shot settings is generally marginal across almost all models. This suggests that while large models are known for their strong in-context learning capabilities (Dong et al., 2022), the complexity and variability of covert advertisements make it difficult for them to effectively utilize the few-shot paradigm and extract meaningful patterns.

5.2 Other Methods

Besides MLMs, we also consider other multimodal methods from related fields to detect covert advertisements as we discussed in Section 2.

We summarize the key differences of these methods in Table 3 and benchmark them in our dataset for the covert advertisement detection task. Notably, while LDSF was originally designed as a tri-modal approach, we adapted it to a bimodal format (image and text) due to the absence of audio information in our dataset.

The experimental results are presented in the lower part of Table 2, indicating that these methods outperform zero-shot or few-shot multimodal large models and exhibit performance comparable to that of fine-tuned models. Among these, MSD displays slightly lower performance, which we attribute to the limitations inherent in using Bi-LSTM compared to the more powerful BERT and CLIP architectures. Further analysis revealed an interesting finding: LDSF and the other three methods exhibited distinct prediction trends for certain samples.

One possible reason is that these positive samples have different representations. According to (Tang et al., 2023), early fusion performs better when modalities are highly correlated, while late fusion excels when features to be identified are primarily in one modality. Covert advertisements may be multiple modalities that are interrelated and echo each other, or they may be presented in only one modality. As a result, different methods yield varying performances depending on the specific sample characteristics.

Table 3: Comparison of other methods: MSD (Cai et al., 2019); HMCAN (Qian et al., 2021); CLIP-GCN (Zhou et al., 2024) and LDSF (Ding et al., 2022). Text denotes the text feature extraction model; Image denotes the image feature extraction model. Multi-modal fusion indicates the fusion phase according to (Gadzicki et al., 2020).

Method	Text	Image	Multi-modal Fusion
MSD	Bi-LSTM	ResNet	Early Fusion
HMCAN	Bert	ResNet	Early Fusion
CLIP-GCN	Clip	Clip	Early Fusion
LDSF	Bert	ResNet	Late Fusion

6 SCAN: Social Media Covert Advertisement Detection using Multi-view Network

Building upon the insights from Section 5.2, our SCAN framework employs a cooperative training mechanism that combines early fusion and late fusion strategies. The framework is structured into two complementary views: a single-channel view and a multi-channel view, as depicted in Figure 3. The single-channel view extracts and processes the individual channel, such as images, text, and comments. The multi-channel view employs multi-head attention to capture cross-channel interactions. After obtaining logits from both views, gated fusion is applied to learn optimal weights and produce the final output. We apply the cooperative training strategy to ensure that these views are trained in a complementary manner, aligning their outputs to promote logit consistency and improve overall performance.

6.1 Notation

For the task of social media covert advertisement detection, we are provided a multimodal dataset $D = \{((I_i, T_i, C_i), y_i) \mid i = 1, \ldots, N\}$, where I_i, T_i, and C_i represent the image, text content, and comments of the i-th instance, respectively, and $y_i \in \mathcal{Y} = \{0, 1\}$ is the binary ground-truth label indicating whether the instance is an advertisement. The objective is to design a multimodal network $F(\cdot)$ that learns feature representations $\mathbf{z}_i$ for each instance and classifies it into one of the two cate-

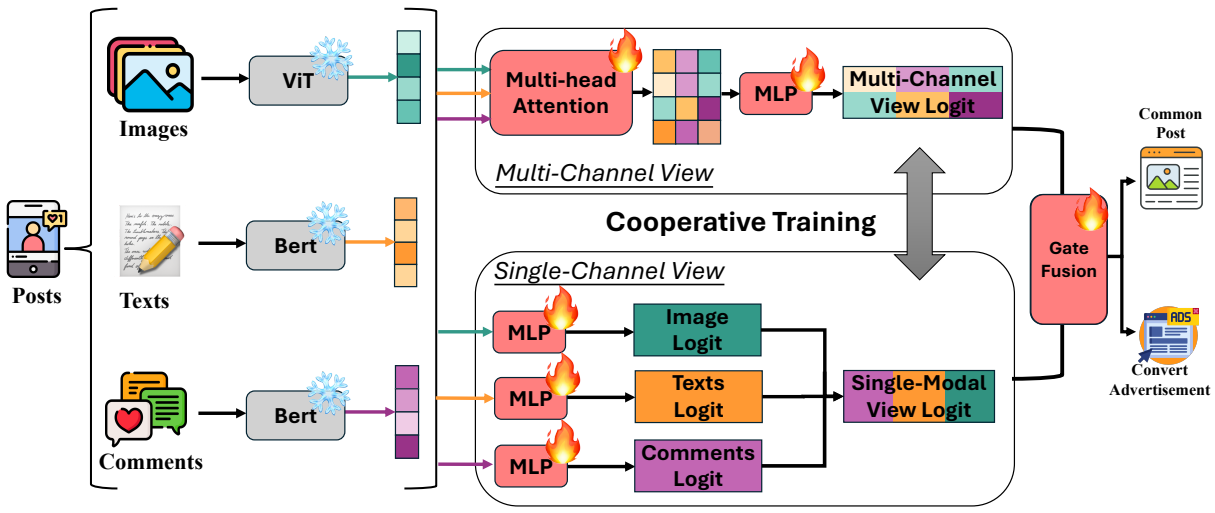

Figure 3: Workflow of SCAN

gories in $\mathcal{Y}$. Ground-truth labels y_i are available only during the training and validation phases.

6.2 Single-Channel View

In the single-channel view, features are independently extracted from each modality to form distinct representations. For images, we use a Vision Transformer (ViT) to process the input image I and obtain the feature vector H_I. Text and comments are processed through a pre-trained BERT model to generate feature vectors H_T and H_C, respectively. Each feature vector is passed through an independent multi-layer perceptron (MLP) to produce logit $f_I(H_I)$, $f_T(H_T)$, and $f_C(H_C)$. The logit from each modality is aggregated to form a fused logit $Z_{single} = \sum_{x \in \{I,T,C\}} f_x(H_x)$.

6.3 Multi-Channel View

To capture interactions across modalities, the unimodal features H_I, H_T, H_C are combined to form a joint representation Z_{joint} through a fusion function $H_{\text{joint}} = \mathcal{F}(H_I, H_T, H_C)$. Here, $\mathcal{F}$ is implemented as a multi-head attention mechanism to focus on complementary information across modalities. Finally, H_{joint} is passed through an MLP to produce the multi-channel view logit over classes: $Z_{\text{joint}} = f_{\text{joint}}(H_{\text{joint}})$.

6.4 Gate Fusion

The Gate Fusion mechanism adaptively integrates outputs from the single-channel and multi-channel views by learning gating weights through a trainable network. Specifically, the logits Z_{single} and Z_{joint} are concatenated and passed through an MLP to produce gating weights G_{single} and G_{joint}. The fused logits are computed as $Z_{\text{final}} = G_{\text{single}} \cdot Z_{\text{single}} + G_{\text{joint}} \cdot Z_{\text{joint}}$, enabling the model to balance contributions from each view dynamically. Finally, the predicted label $\hat{y}$ is obtained by selecting the class with the highest probability: $\hat{y} = \text{argmax } Z_{\text{final}}$.

6.5 Cooperative Training

In the SCAN framework, cooperative learning aligns the single-channel and multi-channel views by leveraging their complementary strengths: the single-channel view captures detailed intra-modal features, while the multi-channel view models cross-modal interactions. This is achieved through a cooperative loss function that combines the final prediction loss with a regularization term measuring the distance between the logits Z_{single} and Z_{joint} of the single-channel and multi-channel views. The loss is defined as:

$$\mathcal{L}_{\text{coop}} = \mathcal{L}_{\text{final}}(P_{\text{final}}, y) + \lambda \cdot D(Z_{\text{single}}, Z_{\text{joint}})$$

where $\mathcal{L}_{\text{final}}$ is the cross-entropy loss, P_{final} is obtained by applying the softmax function to Z_{final}, and D is the distance metric, which we use as the mean squared error. The coefficient λ controls the trade-off between aligning the two views and optimizing predictions, allowing the model to leverage the strengths of both views for more accurate and reliable results.

7 Experiments of SCAN

7.1 Setting

We use the same dataset and metric as described in Section 5. We used a batch size of 16 and

Table 4: The result of SCAN evaluated on Rednote dataset where Qwen2 refers to Qwen2-VL-7B-Instruct here. We choose the top 3 best methods from Section 5 as baselines.

Rednote	ACC ↑	PRE ↑	REC ↑	F1 ↑
Qwen2 (FT)	0.798	0.873	0.697	0.776
HMCAN	0.827	0.789	0.783	0.786
LDSF	0.800	0.780	0.836	0.807
SCAN	**0.877**	**0.875**	**0.879**	**0.877**

Table 5: The result of SCAN evaluated on Instagram dataset where Qwen2 refers to Qwen2-VL-7B-Instruct here and DeepSeek refers to DeepSeek-VL-7B-chat. We choose the top 3 best methods from Section 5 as baselines.

Instagram	ACC ↑	PRE ↑	REC ↑	F1 ↑
DeepSeek (FT)	0.804	0.807	0.800	0.804
CLIP-GCN	0.810	0.815	0.800	0.807
Qwen2 (FT)	0.836	0.794	**0.909**	0.848
SCAN	**0.874**	**0.870**	0.879	**0.874**

trained for 15 epochs. The loss coefficient λ is 1.0. Learning rates are 5×10^{-5} for gate fusion and single-channel views, and 1×10^{-4} for multi-channel views. We used the AdamW optimizer.

7.2 Results

The results are shown in Tables 4 and 5. SCAN achieves outstanding performance on the Rednote dataset, with significant improvements across all metrics, including an F1 score of 0.877. Similarly, on the Instagram dataset, SCAN outperforms existing methods, achieving an F1 score of 0.874 and demonstrating consistent enhancements in accuracy, precision, and recall. These results highlight the effectiveness of the SCAN framework in leveraging both single-channel and multi-channel views for robust performance in detecting covert advertisements across different social media platforms.

7.3 Ablation Studies

We use the Rednote dataset to conduct our ablation studies. Firstly, we tested the effectiveness of each component of SCAN by comparing three baseline models in the Rednote dataset: (1) Only use the Single-channel view; (2) Only use the Multi-channel view; (3) Multi-view model without cooperative training. The results are shown in Figure 4. The results show that the Multi-view model provides a small improvement over the single-channel and multi-channel models, but the performance gain is modest. However, introducing cooperative training in the Multi-view leads to a significant boost in performance, highlighting the effectiveness of cooperative training in enhancing multi-view learning.

Table 6: The result of ablation studies about different λ.

Rednote	ACC ↑	PRE ↑	REC ↑	F1 ↑
$\lambda = 0.5$	0.858	0.835	**0.893**	0.863
$\lambda = 1.0$	**0.877**	**0.875**	0.879	**0.877**
$\lambda = 1.5$	0.856	0.831	0.893	0.861
$\lambda = 2.0$	0.856	0.805	0.940	0.867

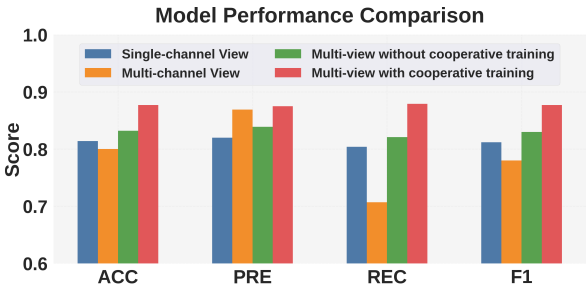

Figure 4: The result of ablation study to examine each component in SCAN.

Secondly, we conducted different coefficient λ values. We try it from 0.1 to 2.0 ($\lambda = 0$ equals to multi-view without cooperative training which we have shown in the first part). The parts of the result are shown in Table 6, and the full result can be found in Table 9. We find that when $\lambda = 1.0$, the SCAN achieves its best performance in all metrics.

8 Conclusion

In this paper, we propose a novel task of covert advertisement detection on social media. To support this task, we provide a manually annotated, challenging, and representative benchmark. We evaluate the performance of several MLMs and related methods on this benchmark. Furthermore, we introduce SCAN, a multi-view cooperative training framework designed to detect covert advertisements on social media effectively. Our method demonstrates strong performance on the benchmark, showcasing its effectiveness and potential for addressing this critical challenge. We believe our work offers a new perspective on social media content moderation.

Limitations

Our study presents a benchmark and framework for social media covert advertisement detection, but we acknowledge certain limitations within our study:

Dataset. Our dataset consists of only two social media platforms and only contains text and image content of posts. If we can have more social media platforms, with video modalities accompanied by users' social relationships, it would be reasonable to conduct research.

Method. Our method uses a fully supervised training approach, so there are a large number of negative samples in our dataset that have not been utilized. We can try to use some semi-supervised learning methods to improve the performance of the method.

Ethic Consideration

The data we collect is completely based on public data platforms and complies with relevant laws and regulations. We also pay attention to data privacy and filter relevant privacy content such as the names of posters and commentators. We ensure that our collection and experimental processes are carried out in a safe and confidential manner.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Michelle A Amazeen. 2023. Native advertising in a mobile era: effects of ability and motivation on recognition in digital news contexts. *Digital Journalism*, 11(6):1130–1153.

S Austin and N Newman. 2015. Attitudes toward sponsored and brand content (native advertising). digital news report 2015. *Reuters Institute for the Study of Journalism/Oxford Internet Institute. http://www. digitalnewsreport. org/essays/2015/attitudes-to-advertising*.

Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. 2020. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38:100311.

Shubham Bharti, Arun Kumar Yadav, Mohit Kumar, and Divakar Yadav. 2022. Cyberbullying detection from tweets using deep learning. *Kybernetes*, 51(9):2695–2711.

Sophie C Boerman, Lotte M Willemsen, and Eva P Van Der Aa. 2017. "this post is sponsored" effects of sponsorship disclosure on persuasion knowledge and electronic word of mouth in the context of facebook. *Journal of Interactive Marketing*, 38(1):82–92.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Glen T Cameron, Kuen-Hee Ju-Pak, and Bong-Hyun Kim. 1996. Advertorials in magazines: Current use and compliance with industry guidelines. *Journalism & Mass Communication Quarterly*, 73(3):722–733.

Colin Campbell and Nathaniel J Evans. 2018. The role of a companion banner and sponsorship transparency in recognizing and evaluating article-style native advertising. *Journal of Interactive Marketing*, 43(1):17–32.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Sheng-wei Tian, and Long Yu. 2022. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Karmen Erjavec. 2004. Beyond advertising and journalism: Hybrid promotional news discourse. *Discourse & Society*, 15(5):553–578.

Yi R Fung and Heng Ji. 2022. A weibo dataset for the 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.05967*.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Yuting Guo, Sudeshna Das, Sahithi Lakamana, and Abeed Sarker. 2023. An aspect-level sentiment analysis dataset for therapies on twitter. *Data in Brief*, 50:109618.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Kevin Lane Keller, MG Parameswaran, and Isaac Jacob. 2010. *Strategic brand management: Building, measuring, and managing brand equity*. Pearson Education India.

Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. 2020. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884.

Seungbae Kim, Jyun-Yu Jiang, and Wei Wang. 2021. Discovering undisclosed paid partnership on social media via aspect-attentive sponsored post learning. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 319–327.

Su Jung Kim, Ewa Maslowska, and Ali Tamaddoni. 2019. The paradox of (dis) trust in sponsorship disclosure: The characteristics and effects of sponsored online consumer reviews. *Decision Support Systems*, 116:114–124.

Susanna Kokko. 2023. Encouraging reading on social media. exploring finnish bookstagram community. Master's thesis.

Szu-Chuang Li, Yu-Ching Chen, Yi-Wen Chen, and Yennun Huang. 2022. Predicting advertisement revenue of social-media-driven content websites: Toward more efficient and sustainable social media posting. *Sustainability*, 14(7):4225.

Song Liang, Ruihang Liu, and Jiansheng Qian. 2021. Fixation prediction for advertising images: Dataset and benchmark. *Journal of Visual Communication and Image Representation*, 81:103356.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Zebo Liu, Kehan Li, Xu Tan, and Jiming Chen. 2020. Iad: A benchmark dataset and a new method for illegal advertising classification. In *ECAI 2020*, pages 2085–2092. IOS Press.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Seema Nagar, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2023. Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining*, 13(1):47.

Yifat Nahmias and Maayan Perel. 2021. The oversight of content moderation by ai: impact assessments and their limitations. *Harv. J. on Legis.*, 58:145.

Michelle R Nelson, Michelle LM Wood, and Hye-Jin Paek. 2009. Increased persuasion knowledge of video news releases: Audience beliefs about news and support for source disclosure. *Journal of Mass Media Ethics*, 24(4):220–237.

Ross D Petty and J Craig Andrews. 2008. Covert marketing unmasked: A legal and regulatory guide for practices that mask marketing messages. *Journal of public policy & marketing*, 27(1):7–18.

Louvins Pierre. 2023. A systematic review of the relationship between covert advertising recognition and consumer attitudes. In *American Academy of Advertising. Conference. Proceedings (Online)*, pages 99–99. American Academy of Advertising.

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

Esther Rozendaal, Moniek Buijzen, and Patti Valkenburg. 2010. Comparing children's and adults' cognitive advertising competences in the netherlands. *Journal of Children and Media*, 4(1):77–89.

Giovanni Santia and Jake Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, pages 531–540.

Prem N Shamdasani, Andrea JS Stanaland, and Juliana Tan. 2001. Location, location, location: Insights for advertising placement on the web. *Journal of Advertising Research*, 41(4):7–21.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. *arXiv preprint arXiv:2203.10885*.

Jinglin Tan. 2024. A critical research on xiaohongshu for information sharing for chinese teenagers. *Profesional de la información*, 33(1).

Qin Tang, Jing Liang, and Fangqi Zhu. 2023. A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*, page 109165.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

Bartosz W Wojdynski and Nathaniel J Evans. 2016. Going native: Effects of disclosure position and language on the recognition and evaluation of online native advertising. *Journal of Advertising*, 45(2):157–168.

Bartosz W Wojdynski and Nathaniel J Evans. 2020. The covert advertising recognition and effects (care) model: Processes of persuasion in native advertising and other masked formats. *International Journal of Advertising*, 39(1):4–31.

Yeqing Yan, Peng Zheng, and Yongjun Wang. 2024. Enhancing large language model capabilities for rumor detection with knowledge-powered prompting. *Engineering Applications of Artificial Intelligence*, 133:108259.

Yufeng Zhou, Aiping Pang, and Guang Yu. 2024. Clipgcn: an adaptive detection model for multimodal emergent fake news domains. *Complex & Intelligent Systems*, pages 1–18.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Assessment Guideline

Observation object:. In order to effectively evaluate whether a post is a hidden advertisement, users should pay comprehensive attention to all parts of the post. Specifically, users need to focus on the image, body content, comments, and even the name of the poster, homepage content, etc.

Identify content:. Users should look for promotions with commercial purposes in the content they observe. For example, product link, brand name, product name, shop name, etc.

Typical examples:. We have summarized several common types of covert advertisements for users' reference, though it is important to note that these are not exhaustive. Covert advertisements can take various forms, such as images displaying the name of the online shop and product, or comments explicitly mentioning the shop name. In some cases, comments may subtly convey product or shop names in complex ways, or images and comments may include product descriptions that hint at where to find the link. Other examples include text making clear references to a product, comments suggesting private messages to share product links, product names visible directly in the image, or even product links hidden in flipped or reversed images. These examples serve as a guide but do not cover all possible manifestations of covert advertisements. We show some examples in Figure 5.

B Dataset

B.1 Post Collection

Rednote. When collecting the Rednote dataset, to ensure its authenticity and diversity, each collector was limited to browsing a maximum of 500 posts per day. This data collection process spanned six days, from 28 September 2024 to 3 October 2024.

As shown in Figure 6, a typical Rednote post consists of five key components. First, images are collected, capturing all visuals associated with the post, which are stored in either JPG or WEBP. Second, the title represents the headline provided by the author to summarize the content. Third, the description encompasses the main body text, offering detailed information or context about the post. Fourth, we collect the data that includes the posting date and associated geographical details, providing temporal and spatial dimensions for analysis. Finally, we gather the top ten comments, reflecting user interactions and engagement with the post, which enrich the dataset's usability for analyzing audience responses.

We had additional analyses of this dataset, and the result of the geographical distribution of the collected posts is in Figure 7. Notably, more than 700 posts were from Guangdong Province, China, and 300 posts were from Zhejiang Province, China. Additionally, more than 120 posts were from the U.S., making it the most represented country outside of China. These geographical details provide

opportunities for users to perform location-based analyses, offering valuable insights into regional trends and user behaviors. Furthermore, we analyzed the posting times of the collected entries, and the results are presented in Figure 8.

Instagram. We also analyzed the distribution of posting times in the Instagram dataset, as illustrated in Figure 9. Compared to the Rednote dataset, Instagram posts exhibit a significantly broader temporal range, with a time span of up to a decade between the earliest and latest posts. Additionally, we examined the distribution of comments and likes within the dataset, which is shown in Figure 10. This analysis provides insights into user engagement trends, revealing the varying levels of interaction across posts over time.

B.2 Post Annotations

To annotate the posts in the benchmark, we hired three Ph.D. students who were well-versed in using both Redote and Instagram. All of them had years of experience with these platforms, which ensured that they were familiar with the typical content, structure, and nuances of posts.

In the preliminary phase, the three annotators labeled 100 samples intuitively based on the definitions summarized in Section 3. Afterward, a feedback session was conducted to review their annotations, analyze consistency and disagreements, and refine the guidelines by addressing ambiguous cases. This iterative process resulted in a consensus-based guideline, and examples of typical covert advertisements, as shown in Figure 5.

It is worth noting that these examples do not represent all types of covert advertisements; they merely illustrate several typical cases. Then, each post was randomly assigned to at least two annotators during the annotation process. In cases where both annotators agreed on the label, no further review was required. However, if there was a disagreement, the third annotator reviewed the post and participated in a discussion to reach a consensus. If all three annotators agreed, the sample was retained; otherwise, it was discarded. Annotators were compensated at a rate of 0.2 USD per sample.

C Benchmark the Baselines

C.1 MLMs

We show the prompts of zero-shot and few-shot as below. It is worth noting that the examples in the few-shot prompt are paired samples extracted from the training set. We also experimented with different numbers of examples. The result is shown in Table 7. The results show that using a single pair leads to poor performance across models, especially with inconsistent precision and recall (e.g., Llava-1.5-13b-hf). Introducing two pairs improves the metrics slightly, but adding more pairs does not consistently enhance performance and leads to instability across datasets and models.

We also show the result of different fine-tuning epochs in Table 8.

D Experiment of SCAN

Ablation studies. We also do an ablation study about the choice of coefficient λ. We try it from 0.1 to 2.0 ($\lambda = 0$ equals to multi-view without cooperative training which we have shown it in Figure 4). The result is shown in Table 6. We find that when $\lambda = 1.0$, the SCAN achieves its best performance in ACC, PRE, REC, or F1.

Zero-shot Prompt

Instruction:
{"role": "system", "content": "You are a helpful assistant. Your task is to analyze images and determine whether they are advertisements. If the image or text is an advertisement, respond with '1'. If the image or text is not an advertisement, respond with '0'. The responses should be concise: '1' for an ad, '0' for non-ad."}
Content:
{"role": "user", "content": "The image is"+image+", and the text is"+ text }
Instruction:
{"role": "system", "content": The responses should be concise: '1' for an ad, '0' for a non-ad. Please respond:}

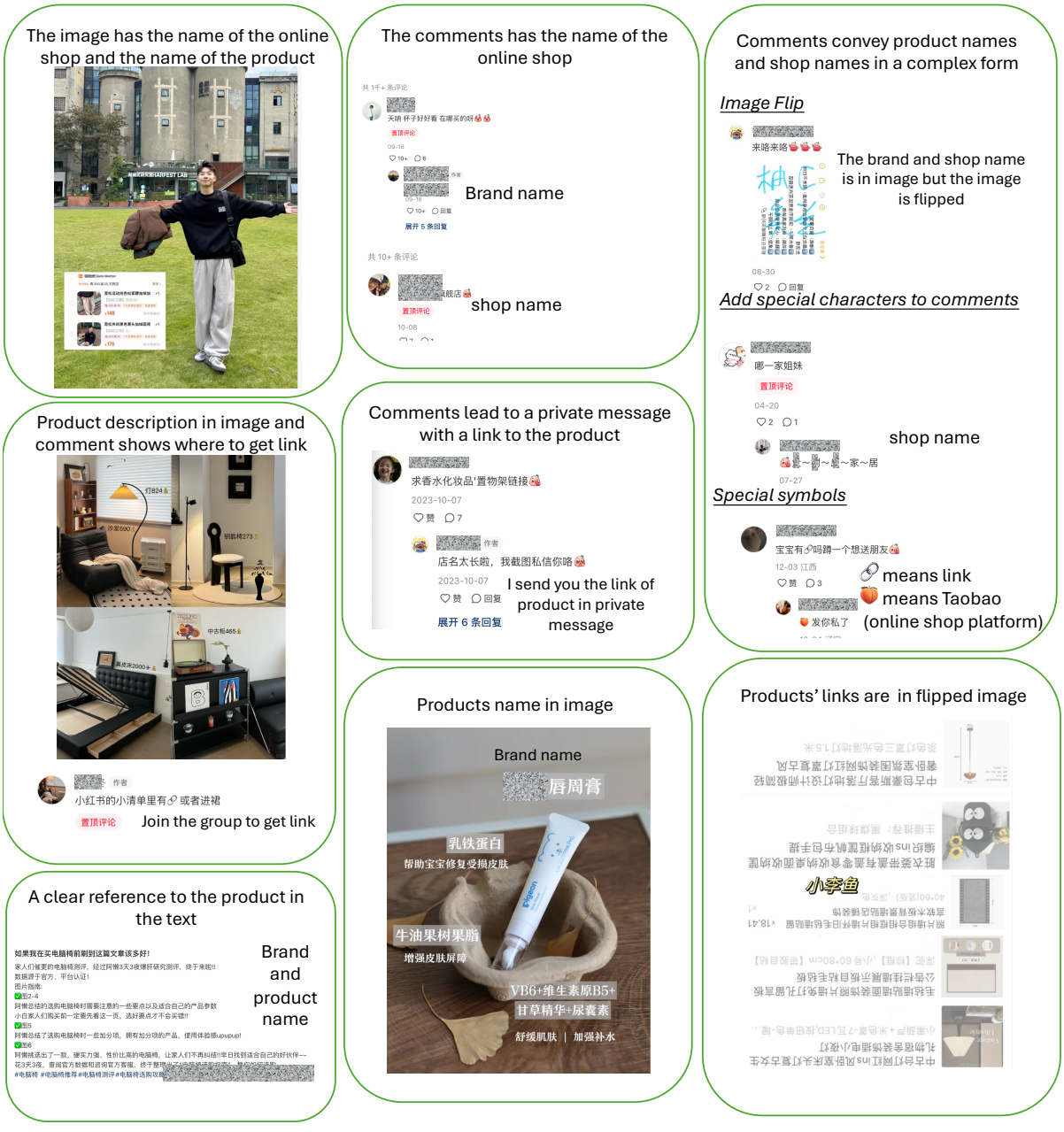

Figure 5: Typical examples of covert advertisements

Few-shot prompt

Instruction:
{"role": "system", "content": "You are a helpful assistant. Your task is to analyze images and determine whether they are advertisements. If the image or text is an advertisement, respond with ’1’. If the image or text is not an advertisement, respond with ’0’. The responses should be concise: ’1’ for an ad, ’0’ for non-ad. "}
Examples: {"role": "user", "content": "The image is"+image+", and the text is"+ text }
{"role": "assistant", "content": "1".}
{"role": "user", "content": "The image is"+image+", and the text is"+ text }
{"role": "assistant", "content": "0".}
.... **Content:**
{"role": "user", "content": "The image is"+image+", and the text is"+ text }
instruction:
{"role": "system", "content": "You are a helpful assistant. Your task is to analyze images and determine whether they are advertisements. If the image or text is an advertisement, respond with ’1’. If the image and text are not an advertisement, respond with ’0’. "}

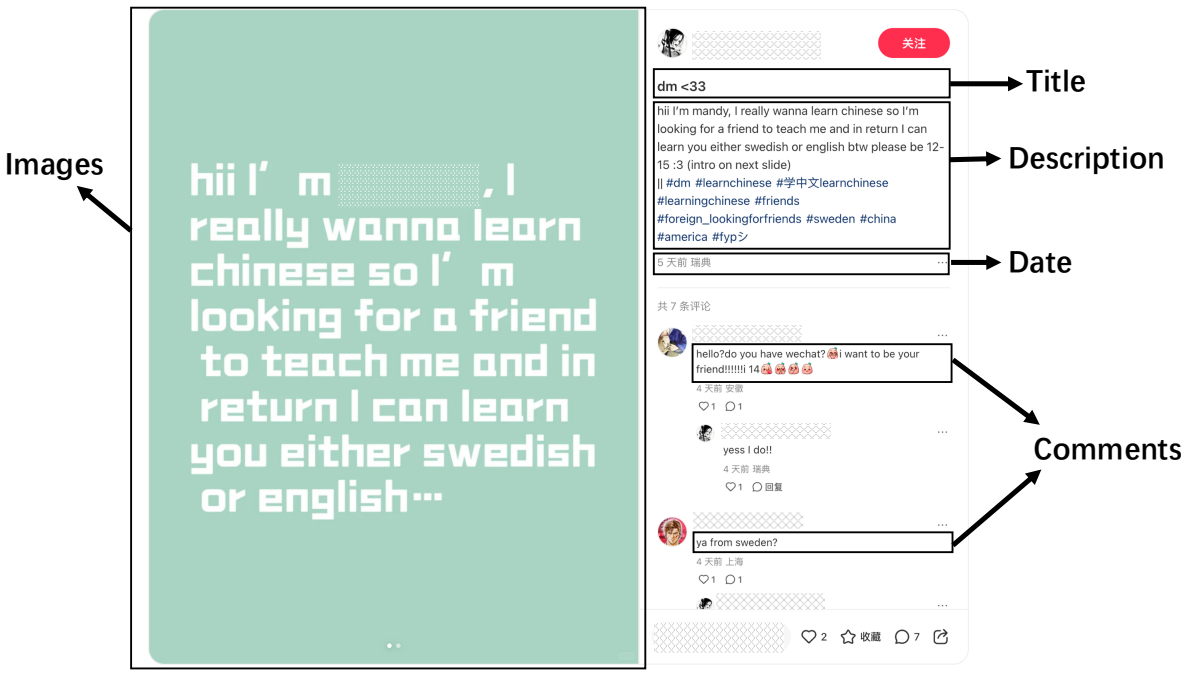

Figure 6: Rednote post collection example which data has been desensitized

Model		Rednote				Instagram			
		ACC ↑	PRE ↑	REC ↑	F1 ↑	ACC ↑	PRE ↑	REC ↑	F1 ↑
Qwen2-VL-7B-Instruct	1	0.586	0.555	0.819	0.662	0.632	0.633	0.627	0.630
	2	0.708	0.712	0.688	0.700	0.650	0.651	0.646	0.648
	3	0.659	0.652	0.682	0.667	0.690	0.783	0.527	0.630
	4	0.709	0.740	0.645	0.689	0.641	0.670	0.555	0.607
	5	0.664	0.632	0.782	0.699	0.610	0.613	0.591	0.602
Llava-1.5-13b-hf	1	0.542	1.000	0.007	0.139	0.514	0.516	0.455	0.483
	2	0.583	0.909	0.184	0.305	0.611	0.797	0.268	0.402
	3	0.600	0.867	0.236	0.371	0.605	0.871	0.246	0.383
	4	0.577	0.904	0.173	0.290	0.618	0.825	0.300	0.440
	5	0.581	0.875	0.191	0.313	0.605	0.829	0.264	0.400
DeepSeek-VL-7B-chat	1	0.623	0.645	0.546	0.591	0.665	0.786	0.426	0.553
	2	0.632	0.630	0.618	0.624	0.670	0.802	0.426	0.557
	3	0.586	0.584	0.600	0.592	0.686	0.836	0.463	0.596
	4	0.623	0.617	0.646	0.631	0.664	0.891	0.373	0.526
	5	0.654	0.685	0.573	0.624	0.700	0.907	0.446	0.598

Table 7: The result of benchmarking the Multi-modal Large model in different few-shot sample pairs.

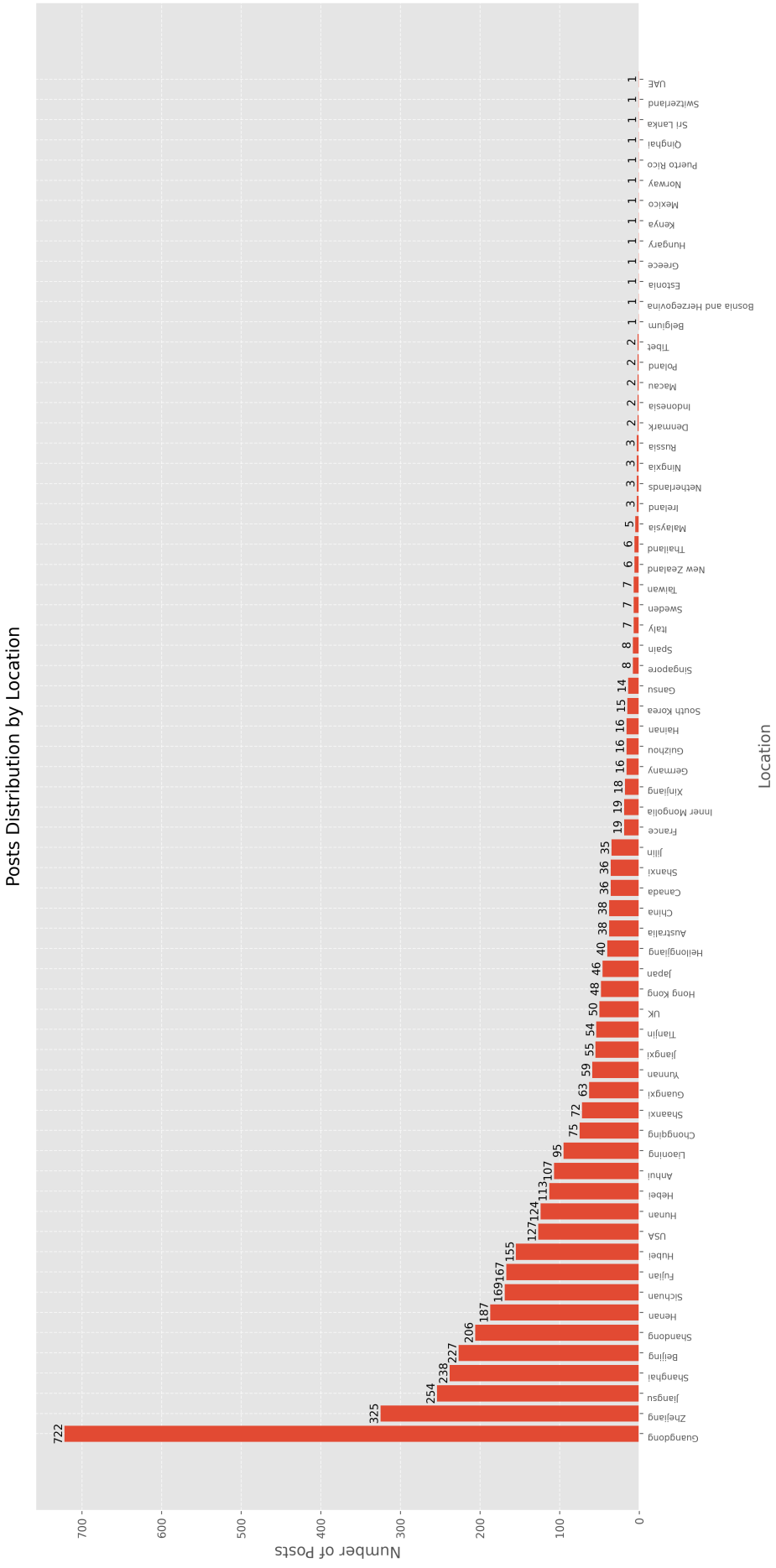

Figure 7: Rednote location distribution

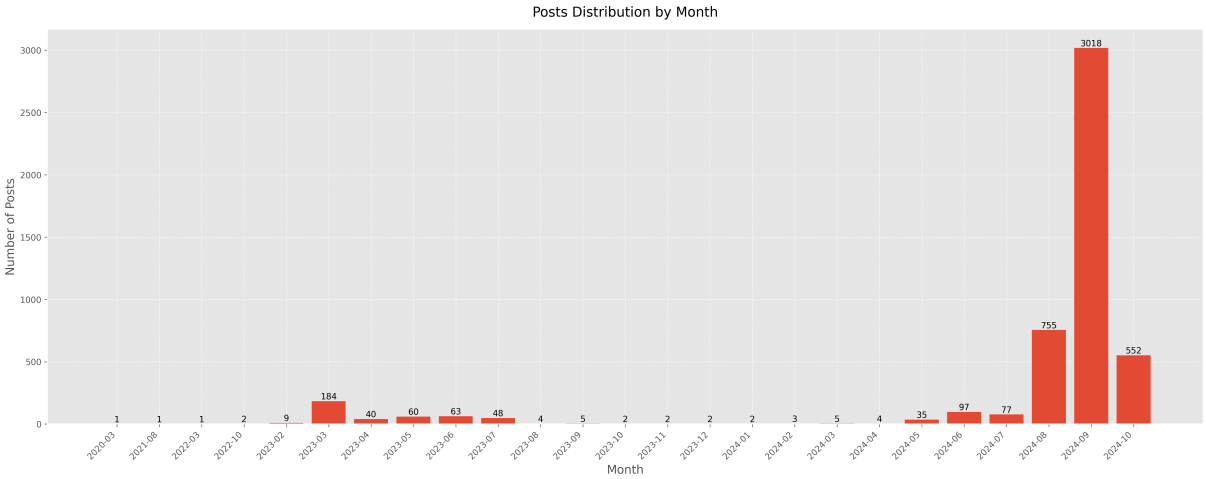

Figure 8: Rednote month distribution

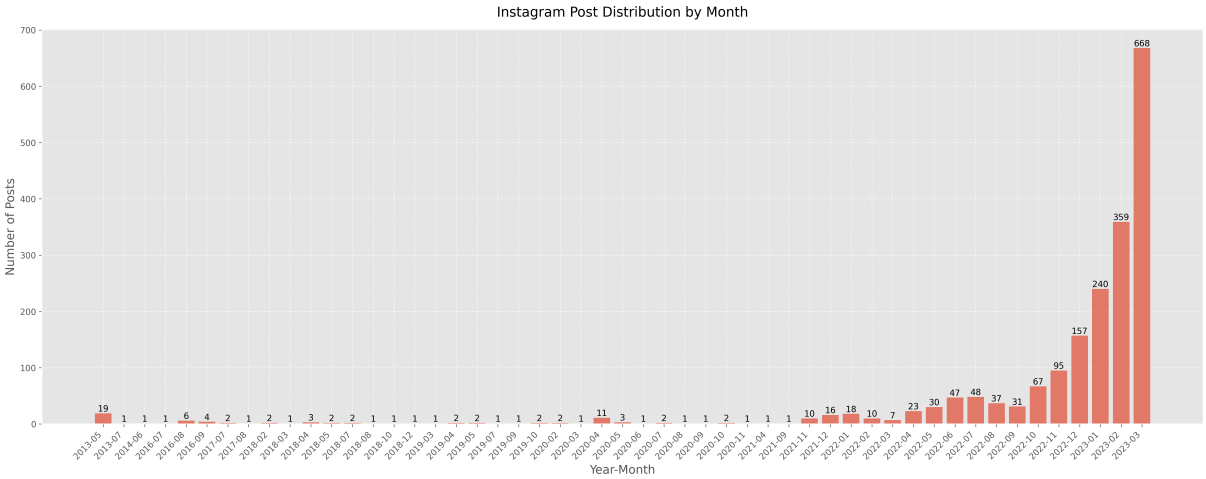

Figure 9: Instagram dataset post time distribution

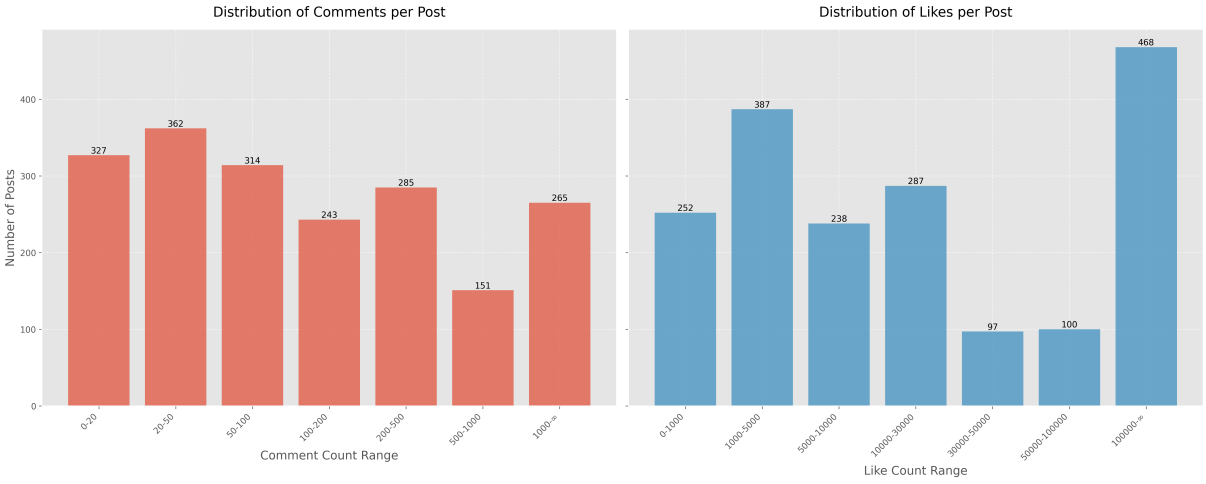

Figure 10: Instagram dataset like and comments number distribution

Model		Rednote				Instagram			
		ACC ↑	PRE ↑	REC ↑	F1 ↑	ACC ↑	PRE ↑	REC ↑	F1 ↑
Qwen2-VL-7B-Instruct	5	0.798	0.873	0.697	0.776	0.836	0.794	0.909	0.848
	10	0.791	0.881	0.673	0.763	0.832	0.829	0.836	0.833
	15	0.809	0.947	0.654	0.774	0.823	0.748	0.972	0.846
Llava-1.5-13b-hf	5	0.778	0.877	0.651	0.747	0.796	0.828	0.746	0.785
	10	0.786	0.943	0.609	0.740	0.804	0.868	0.718	0.786
	15	0.772	0.857	0.654	0.742	0.791	0.827	0.736	0.778
DeepSeek-VL-7B-chat	5	0.782	0.798	0.755	0.776	0.804	0.807	0.800	0.804
	10	0.781	0.793	0.764	0.778	0.804	0.813	0.790	0.802
	15	0.772	0.789	0.746	0.766	0.805	0.845	0.745	0.792

Table 8: The result of benchmarking the Multi-modal Large model in different epochs of fine-tuning.

Rednote	ACC ↑	PRE ↑	REC ↑	F1 ↑
$\lambda = 0.1$	0.846	0.828	0.874	0.851
$\lambda = 0.2$	0.858	0.841	0.884	0.862
$\lambda = 0.3$	0.851	0.833	0.879	0.855
$\lambda = 0.4$	0.860	0.830	0.907	0.867
$\lambda = 0.5$	0.858	0.835	0.893	0.863
$\lambda = 0.6$	0.848	0.798	0.935	0.861
$\lambda = 0.7$	0.856	0.834	0.888	0.860
$\lambda = 0.8$	0.858	0.811	0.935	0.868
$\lambda = 0.9$	0.865	0.828	0.921	0.872
$\lambda = 1.0$	0.877	0.875	0.879	0.877
$\lambda = 1.1$	0.865	0.837	0.907	0.871
$\lambda = 1.2$	0.867	0.829	0.926	0.875
$\lambda = 1.3$	0.861	0.827	0.912	0.867
$\lambda = 1.4$	0.854	0.839	0.874	0.857
$\lambda = 1.5$	0.856	0.831	0.893	0.861
$\lambda = 1.6$	0.858	0.824	0.912	0.865
$\lambda = 1.7$	0.854	0.830	0.888	0.858
$\lambda = 1.8$	0.851	0.803	0.930	0.862
$\lambda = 1.9$	0.854	0.839	0.874	0.857
$\lambda = 2.0$	0.856	0.805	0.940	0.867

Table 9: Ablation studies about different λ