AdDriftBench: A Benchmark for Detecting Data Drift and Label Drift in Short Video Advertising

Anonymous ACL submission

Abstract

With the commercialization of short video platforms (SVPs), the demand for compliance auditing of advertising content has grown rapidly. The rise of large vision-language models (VLMs) offers new opportunities for automating ad content moderation. However, short video advertising scenarios present unique challenges due to **data drift (DD)** and **label drift (LD)**. DD refers to rapid shifts in data distribution caused by advertisers to evade platform review mechanisms. LD arises from the evolving and increasingly standardized review guidelines of SVPs, which effectively alter the classification boundaries over time. Despite the significance of these phenomena, there is currently a lack of benchmark tools designed to evaluate model performance under such conditions. To address this gap, we propose **AdDriftBench (ADB)**. The ADB dataset consists of 3,480 short video ads, including 2,280 examples labeled under data drift scenarios, designed to evaluate the generalization capabilities of VLMs under rapidly shifting content distributions. An additional 1,200 examples represent label drift scenarios, aimed at assessing VLMs' abilities in instruction following and fine-grained semantic understanding under varying auditing standards. Through extensive experiments on 16 open-source VLMs, we find that current models perform moderately in short video advertising contexts, particularly in handling fine-grained semantics and adapting to shifting instructions. Our dataset will be made publicly available.

1 Introduction

The commercialization of short video platforms (SVPs) has led to a growing demand for the moderation of short video advertisements. Traditional content moderation methods rely heavily on manual rules and small-scale models(Szwed et al., 2016; Liu et al., 2020). Recently, vision-language models

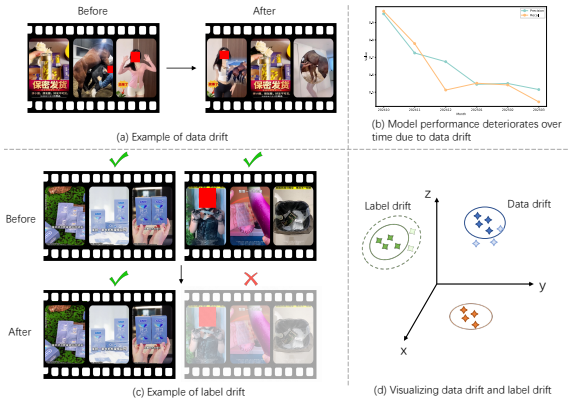


Figure 1: Introduction to DD and LD. (a) DD in the erectile dysfunction drug advertisement scenario. Advertisers use materials with different data distributions to bypass the current review system. (b) Visualization of DD. A multimodal fusion model trained on data before 202410 shows a gradual decline in performance over time. (c) LD in the condom advertisement scenario. After rule tightening, condom advertisements involving vulgar content are rejected. (d) Visualization of DD and LD. The borders represent review rules, and the quadrilaterals represent video distributions.

(VLMs) have demonstrated impressive capabilities in both visual and textual understanding(Wu et al., 2024b; Bai et al., 2025; Zhu et al., 2025), showing strong potential in tasks such as content comprehension and violation detection. However, the short video advertising domain is characterized by large-scale data drift (DD) and label drift (LD), posing new challenges for VLMs in terms of fine-grained semantic understanding and strict instruction following.

DD refers to frequent shifts in data distribution caused by advertisers aggressively modifying their content to evade platform moderation policies. As illustrated in Figure 1(a), whereas previously advertisers might embed explicit violations (e.g., horse mating scenes) directly into videos, they now often overlay such content using picture-in-picture

(PIP) techniques. Figure 1(b) illustrates the performance degradation of a multimodal small model over time, highlighting how DD contributes to the model's decreasing accuracy.

LD refers to changes in the classification boundarie, resulting from the increasingly standardized moderation rules on SVPs. As illustrated in Figure 1(c), condom advertisements with suggestive content were previously allowed but are now considered violations under stricter policies. Figure 1(d) visualizes both DD and LD effects: quadrilaterals represent the distribution of video ads, and circles represent the classification boundaries.

Several benchmarks(Chen et al., 2024b; Xu et al., 2025; Lu et al., 2025) have been proposed to evaluate video content compliance (see Table 1). However, none of them simultaneously consider both DD and LD. To fill this gap, we propose AdDrift-Bench (ADB)—a new benchmark specifically designed for short video advertising scenarios. ADB consists of 3,480 video ads, including 2,280 DD samples spanning 6 primary risk categories and 12 secondary categories. These samples are temporally segmented to assess VLMs' generalization under drastic distribution shifts. The LD portion includes 1,200 samples covering 10 primary risk categories, where each video is evaluated under two audit standards (e.g., "lenient" vs. "strict" prompts) to test the VLMs' instruction-following and fine-grained semantic understanding abilities.

To ensure data quality, we applied similarity-based deduplication, used models to pre-filter high-risk cases, and involved professional human reviewers for final validation. Through a comparison of 16 widely-used open-source VLMs, we find that their compliance identification performance in short video advertising scenarios is suboptimal. This highlights the need for improvement in both instruction following and fine-grained semantic reasoning. Our contributions are as follows:

- We identify and formalize the challenges of data drift and label drift in short video advertising.
- We introduce AdDriftBench (ADB), the first benchmark designed to evaluate VLMs' robustness to both data and label drift in short video ad scenarios.
- We conduct comprehensive comparative and ablation studies on 16 open-source VLMs, drawing eight key findings that offer valuable insights for future research.

Benchmarks	SV	Ad	DD	LD
SafeWatch(Chen et al., 2024b)	✗	✗	✔	✗
MMDT(Xu et al., 2025)	✗	✗	✔	✗
XD-Violence(Wu et al., 2020)	✗	✗	✗	✗
UCF-Crime(Sultani et al., 2018)	✗	✗	✗	✗
FakeSV(Qi et al., 2023)	✔	✗	✔	✗
FVC(Papadopoulou et al., 2019)	✔	✗	✔	✔
LSPD(Phan et al., 2022)	✗	✗	✗	✗
KuaiMod(Lu et al., 2025)	✔	✗	✗	✔

Table 1: Comparison of the dimensions involved in different benchmarks. SV represents Short Videos, Ad represents Advertisement scenarios, and DD and LD represent Data Drift and Label Drift, respectively.

2 Related Work

2.1 VLMs and Evaluations

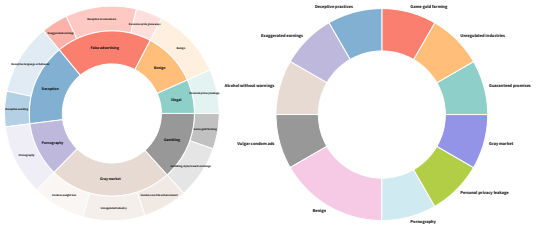
In recent years, VLMs have made significant advancements. DeepSeek-VL2(Wu et al., 2024b) utilizes a Mixture-of-Experts (MoE) architecture, achieving outstanding performance across multiple multimodal benchmarks. Qwen2.5-VL(Bai et al., 2025), consisting of a visual encoder and a language model, supports dynamic resolution and frame rate training. Qwen2-VL(Wang et al., 2024b) processes images of arbitrary resolution, employing 2D-RoPE position encoding to replace traditional absolute position encoding, thereby better capturing the two-dimensional positional information of images. InternVL2.5(Chen et al., 2024c) enhances the model's inference capabilities and multi-image information integration by incorporating additional multi-image datasets. LLaVA-OneVision(Li et al., 2024a), built on the LLaVA architecture, exhibits strong cross-modal transfer capabilities. LLaVA-NeXT-Video(Liu et al., 2024a), based on LLaVA-NeXT, improves video understanding through supervised fine-tuning (SFT) and direct preference optimization (DPO) on video data.

VLMs are typically validated on various public benchmarks to assess their general visual understanding and generation capabilities. Benchmarks such as MMBench(Liu et al., 2024b), MM-Star(Chen et al., 2024a), MuirBench(Wang et al., 2024a), BLINK(Fu et al., 2024b), CRPE(Wang et al., 2024c), and HallBench(Guan et al., 2024) design general VQA tasks to evaluate VLMs' general visual understanding ability. AI2D(Kembhavi et al., 2016), TextVQA(Singh et al., 2019), DocVQA(Mathew et al., 2021), and InfoVQA(Mathew et al., 2022) focus on

Question: You are a short video advertisement reviewer, ···, the definition of primary violation scenes is as follows, ···, the definition of secondary violation scenes is as follows, ···, and output in JSON format, such as: {"Primary Scene" : "Deception", "Secondary Scene": "Deceptive wording"}
Ground Truth: {"Primary Scene" : "Deception", "Secondary Scene": "Deceptive wording"}

InternVL3-9B: {"Primary Scene" : "Deception", "Secondary Scene": "Deceptive wording"}
Qwen2.5-VL-7B: {"Primary Scene" : " False advertising ", "Secondary Scene": " Deceptive to consumers "}
Qwen2.5-VL-32B: : {"Primary Scene" : "Deception", "Secondary Scene": "Deceptive wording"}
Qwen2.5-VL-72B: : {"Primary Scene" : "Deception", "Secondary Scene": "Deceptive wording"}

Figure 2: AdDriftBench example sampling and model output.



(a) Data drift scene distribution (b) Label drift scene distribution

Figure 3: Distribution of data drift and dabel drift scenarios. (a) DD covers 7 primary scenarios and 13 secondary scenarios (including the benign scenario). LD includes 11 main scenarios (including the benign scenario).

evaluating VLMs' document understanding and OCR capabilities. Some benchmarks, like CountBench(Paiss et al., 2023), specifically assess VLMs' spatial understanding abilities. Video-MME(Fu et al., 2024a), Video-MMMU(Hu et al., 2025), MMVU(Zhao et al., 2025), MVBench(Li et al., 2024b) and LongVideoBench(Wu et al., 2024a) focus on evaluating VLMs' multimodal understanding abilities in the domains of video understanding and grounding.

2.2 Multimodal Safety-Related Benchmarks

Currently, the academic community has proposed various benchmark datasets focused on image-video safety, which can be broadly categorized into two types: general safety capability evaluation and single-scene safety capability evaluation. For general safety scenarios, MMDT(Xu et al., 2025) has introduced a comprehensive safety evaluation platform for VLMs, covering six key dimensions: security, hallucination, bias and fairness, privacy, adversarial robustness, and OOD generalization. SafeWatch-Bench(Chen et al., 2024b) focuses on video content safety and has built an ultra-large dataset containing 2 million videos. KuaiMod(Lu et al., 2025) is the first benchmark proposed by KuaiShou for general safety scenarios in short videos. However, it focuses on generic content and addresses only the issue of label drift. In contrast, our proposed ADB benchmark targets compliance-related violations in short video advertising scenarios and explicitly tackles both data drift and label drift.

For single safety scenarios, FakeSV(Qi et al., 2023) focuses on short video fake news detection, emphasizing the integration of multimodal cues and social context. LSPD(Phan et al., 2022) provides large-scale benchmarks for multi-granularity harm detection. XD-Violence(Wu et al., 2020) targets violence scene detection, while UCF-Crime(Sultani et al., 2018) focuses on abnormal behavior detection, covering 13 types of abnormal events. FVC-2018(Papadopoulou et al., 2019) focuses on fake news videos, used for multi-version consistency verification and rumor tracing research.

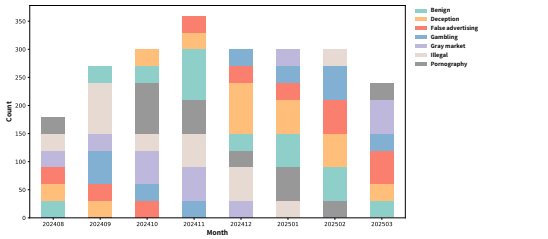
However, the above datasets do not adequately address the common issues of data drift and label drift present in short video advertising scenarios. Frequent material iteration by advertisers causes drastic changes in data distribution over time, while the continuous refinement of platform review rules leads to the dynamic evolution of labels. Both of these factors inherently raise the demands on VLMs' OOD generalization ability and fine-grained semantic understanding and instruction-following capabilities. To fill this gap, we propose AdDriftBench—the first multimodal safety benchmark for short video advertisements. This dataset explicitly constructs evaluation tracks based on data drift and label drift to systematically test the robustness and transferability of VLMs in real-world advertising review scenarios.

3 AdDriftBenchmark

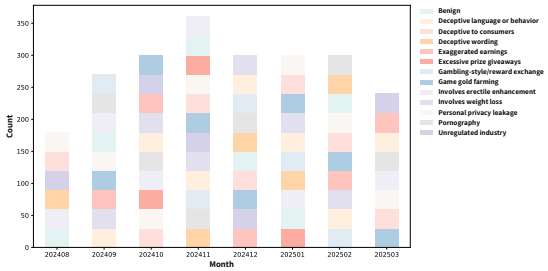
3.1 Task Design

Data source. AdDriftBench (ADB) focuses on evaluating the ability of VLMs to handle data drift and label drift in short video advertising scenarios. To this end, we selected 30,000 short video advertisements from the Kuaishou platform. After filtering, there are 2,280 short videos in the data drift scenario and 1,200 short videos in the label drift scenario, an example is illustrated in Figure 2.

To ensure the results are convincing, we aimed to maintain a balanced distribution of data across each scenario, with the specific scene distribution shown in Figure 3. For the data drift scenario, we selected 6 primary scenes (such as gray market,



(a) Monthly distribution of primary scenes



(b) Monthly distribution of secondary scenes

Figure 4: Data drift monthly distribution. We ensure that each secondary scenario contains 30 videos per month. If a given secondary scenario has fewer than 30 samples in a month, it is removed for that month.

pornography, gambling, and false advertising) and 12 secondary scenes (such as involves weight loss, unregulated industry, and personal privacy leakage). For the label drift scenario, we selected 10 scenes (such as guaranteed promises, vulgar condom ads, and gray market), with detailed scene definitions provided in Appendix A (Table 4 and 5).

Data drift scenario design. In the short video advertising scenario, advertisers are continuously iterating materials in an attempt to bypass platform review rules, leading to the same violation scene appearing with different distributions of violating videos. We trained a 7-class small model and observed that over time, both the precision and recall of the model decreased (as illustrated in Figure 1(b)). Since small models tend to overfit the data distribution of the training set, it can be assumed that the data distribution changes each month. Therefore, we selected data from 8 months, ensuring that each secondary scene appeared with at least 30 samples per month. The distribution of primary and secondary scenes for each month is illustrated in the Figure 4. We evaluate VLMs' ability to handle data drift by measuring their precision and recall in different scenes across different months.

Label drift scenario design. As illustrated in Figure 1(c), in short video advertising, label drift arises as platforms become more regulated and review policies grow increasingly strict. Videos that previously passed review may now be rejected. To study this, we selected 10 scenarios where rule tightening has caused label drift (as illustrated in Figure 3(b)), and evaluated VLMs' robustness to label drift by prompting them with different review criteria. Specifically, we design two sets of prompts—lenient and strict—aimed at assessing the VLMs' instruction-following ability and fine-grained semantic understanding.

3.2 Dataset Curation

The data collection process includes three parts: similarity-based deduplication, multimodal small model screening, and manual review, as illustrated in Figure 5. The details of each part will be described next. All video data has been anonymized.

Similarity-based deduplication. We downloaded 30,000 short video advertisement data from the Kuaishou platform and first performed similarity-based deduplication (as illustrated in Figure 6). Since advertisers often upload similar advertisement materials repeatedly to gain exposure at a low cost, we need to perform inter-video similarity deduplication (Figure 6(a)). Additionally, we extracted dense frames from each video at a rate of 1 frame per second. There are many similar frames within the same video, so we also need to perform intra-video deduplication (Figure 6(b)). The distribution of video frames before and after intra-video deduplication is illustrated in Appendix D (Figure 15).

Specifically, for inter-video deduplication (Figure 6(a)), we used VIT-B-32 of CLIP(Radford et al., 2021) to extract the embedding for each frame and averaged the embeddings of all frames in the same video to obtain the global feature for the video. We computed the global features for all videos and calculated the cosine similarity. Videos with a similarity threshold greater than 0.92 were grouped into a connected subgraph, and only one node (one video) was retained for each connected subgraph, reducing the video count from 30,000 to approximately 24,000. Similarly, for intra-video deduplication (Figure 6(b)), we treated each frame's embedding as a node in the connected subgraph and retained only one node (one frame) for each connected subgraph.

Intra-video deduplication is mainly performed to improve VLMs' performance while saving inference costs. Since the current VLMs have a limited context window that cannot accommodate all video frames, we aim to reduce frame-level redundancy

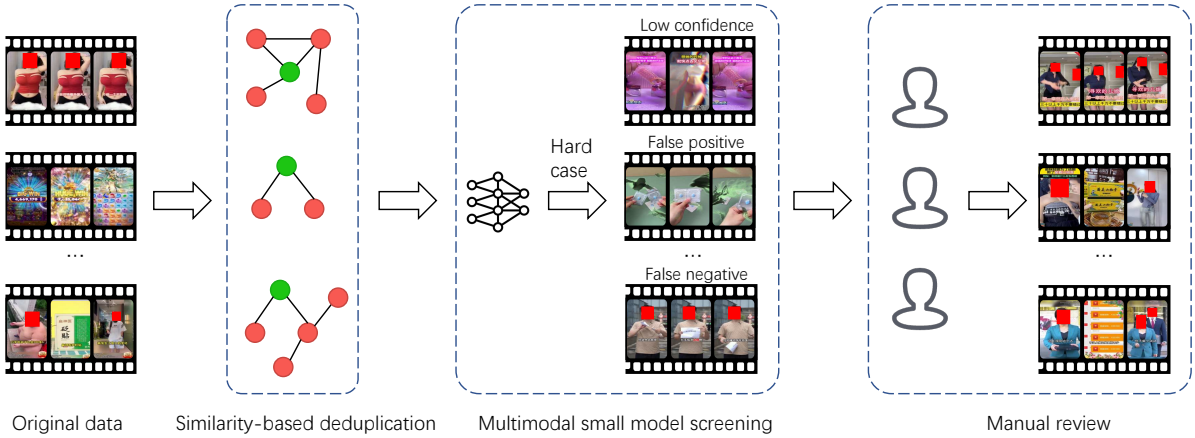


Figure 5: Data collection pipeline. We adopted a three-step process—similarity-based deduplication, multimodal small model screening, and manual review—to ensure model quality and complexity.

to maximize the utilization of input tokens.

Multimodal small model screening for hard cases. We trained a seven-class multimodal small model based on the data drift scenarios shown in Figure 3(a); detailed model architecture and settings are provided in the Appendix B (Table 10). To increase the difficulty of the dataset, we selected 5,000 videos from the model's predictions that included low-confidence samples, false positives, and false negatives. To ensure balanced distribution across both data drift and label drift scenarios, we ultimately sampled a total of 3,480 videos—2,280 for data drift and 1,200 for label drift.

Manual review. To ensure the quality of ADB dataset, all 3,480 hard cases were manually reviewed by a team of six professionally trained short video reviewers. Prior to the review process, we confirmed that all reviewers had a clear understanding of the review guidelines (Table 4 and 5).

4 Experiments

4.1 Experimental Setup

Model Configurations. We evaluated ADB on 16 mainstream open-source models, including the DeepSeek-VL2 series(Wu et al., 2024b), InternVL2.5 series(Chen et al., 2024c), InternVL3 series(Zhu et al., 2025), Qwen2-VL series(Wang et al., 2024b), Qwen2.5-VL series(Bai et al., 2025), LLaVA-NeXT-Video-7B(Liu et al., 2024a), and LLaVA-OneVision-7B(Li et al., 2024a). Detailed model configurations are provided in Appendix Table 11. All experiments were conducted on two H20 GPUs. The detailed configurations of the VLMs we evaluated are provided in Appendix F (Table 11). The input prompt is provided in Appendix G (Figure 16, 17, 19, 18).

Figure 6: Similarity-based deduplication. Nodes (videos or frames) with high similarity are grouped into a connected subgraph, and only one node is retained from each connected subgraph. The purpose of deduplication is to reduce the inference cost of VLMs.

Evaluation Metrics. For data drift scenarios, we computed precision $\mathcal{P}$, recall $\mathcal{R}$, and $\mathcal{F}_1$ for each month and each scene. We used the average $\overline{\mathcal{P}}$, $\overline{\mathcal{R}}$, and $\overline{\mathcal{F}_1}$ across months to evaluate the model's risk identification capability in short video advertising. Additionally, we calculated the ratio of the standard deviation to the mean ($\mathcal{SDM}$) for $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ across all months (as illustrated in Equation 1) to assess the model's robustness to data drift.

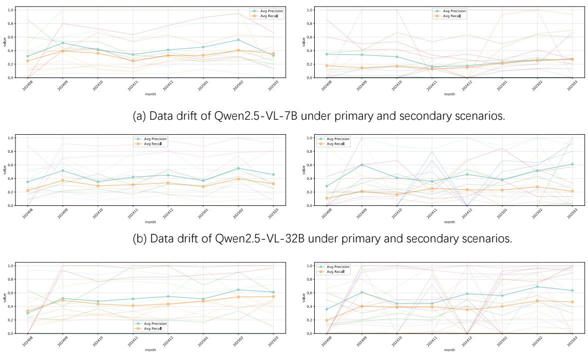
$$\mathcal{SDM}_{\mathcal{P}} = \frac{\sigma_{\mathcal{P}}}{\overline{\mathcal{P}}}, \quad \mathcal{SDM}_{\mathcal{R}} = \frac{\sigma_{\mathcal{R}}}{\overline{\mathcal{R}}}, \quad \mathcal{SDM}_{\mathcal{F}_1} = \frac{\sigma_{\mathcal{F}_1}}{\overline{\mathcal{F}_1}} \quad (1)$$

where $\sigma_{\mathcal{P}}$, $\sigma_{\mathcal{R}}$, $\sigma_{\mathcal{F}_1}$ denotes the standard deviation and $\overline{\mathcal{P}}$, $\overline{\mathcal{R}}$, $\overline{\mathcal{F}_1}$ denotes the mean. A lower $\mathcal{SDM}$ indicates better adaptability to distribution shifts.

For label drift scenarios, we computed and compared the average $\overline{\mathcal{P}}$, $\overline{\mathcal{R}}$, and $\overline{\mathcal{F}_1}$ before and after the drift to evaluate the model's ability to handle label drift.

	202408			202409			202410			202411			202412			202501			202502			202503			Avg			SDM		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\overline{\mathcal{P}}$	$\overline{\mathcal{R}}$	$\overline{\mathcal{F}_1}$	$\overline{\mathcal{P}}\downarrow$	$\overline{\mathcal{R}}\downarrow$	$\overline{\mathcal{F}_1}\downarrow$
ds-vl2-tiny	0.10	0.09	0.03	0.02	0.06	0.02	0.00	0.04	0.01	0.02	0.13	0.04	0.11	0.16	0.06	0.07	0.15	0.05	0.10	0.15	0.05	0.01	0.06	0.02	0.05	0.11	0.04	0.86	0.46	0.51
ds-vl2-small	0.08	0.11	0.07	0.18	0.08	0.04	0.20	0.07	0.04	0.07	0.12	0.06	0.20	0.13	0.07	0.15	0.12	0.07	0.24	0.14	0.10	0.08	0.08	0.06	0.15	0.11	0.06	0.44	0.25	0.30
ds-vl2	0.21	0.13	0.10	0.12	0.17	0.13	0.11	0.14	0.11	0.15	0.18	0.14	0.11	0.15	0.10	0.14	0.16	0.13	0.20	0.22	0.18	0.16	0.15	0.10	0.15	0.16	0.12	0.26	0.17	0.22
internvl2.5-2b	0.08	0.17	0.09	0.17	0.16	0.09	0.24	0.14	0.06	0.24	0.15	0.10	0.30	0.12	0.06	0.18	0.16	0.09	0.24	0.19	0.11	0.11	0.20	0.12	0.20	0.16	0.09	0.38	0.16	0.24
internvl2.5-4b	0.18	0.14	0.11	0.25	0.24	0.22	0.15	0.18	0.14	0.24	0.23	0.20	0.23	0.16	0.14	0.29	0.26	0.23	0.30	0.28	0.24	0.31	0.28	0.25	0.24	0.22	0.19	0.23	0.25	0.28
internvl2.5-8b	0.30	0.20	0.18	0.42	0.28	0.24	0.38	0.26	0.20	0.33	0.26	0.20	0.38	0.18	0.16	0.35	0.23	0.19	0.32	0.29	0.26	0.47	0.24	0.20	0.37	0.24	0.20	0.15	0.16	**0.16**
internvl3-2b	0.04	0.11	0.06	0.15	0.10	0.08	0.10	0.08	0.06	0.14	0.14	0.10	0.12	0.14	0.10	0.22	0.15	0.13	0.19	0.16	0.14	0.20	0.11	0.08	0.15	0.12	0.09	0.41	0.22	0.32
internvl3-9b	0.16	0.11	0.08	0.58	0.26	0.30	0.43	0.23	0.26	0.39	0.28	0.26	0.41	0.31	0.28	0.45	0.29	0.26	0.50	0.34	0.32	0.38	0.16	0.18	0.41	0.25	0.24	0.29	0.31	0.32
llava-nextvideo-7b	0.14	0.10	0.11	0.16	0.07	0.08	0.22	0.03	0.03	0.08	0.03	0.02	0.06	0.03	0.03	0.20	0.12	0.12	0.05	0.05	0.04	0.37	0.12	0.14	0.16	0.07	0.07	0.66	0.58	0.66
llava-onevision-7b	0.03	0.16	0.04	0.06	0.16	0.08	0.06	0.14	0.06	0.10	0.16	0.08	0.08	0.18	0.09	0.13	0.20	0.12	0.16	0.20	0.12	0.02	0.12	0.03	0.08	0.16	0.08	0.6	0.17	0.43
qwen2-vl-2b	0.08	0.11	0.08	0.14	0.10	0.08	0.10	0.07	0.04	0.15	0.08	0.07	0.12	0.10	0.09	0.13	0.08	0.08	0.23	0.12	0.12	0.08	0.10	0.08	0.13	0.10	0.08	0.38	0.18	0.28
qwen2-vl-7b	0.22	0.16	0.15	0.18	0.14	0.13	0.12	0.14	0.11	0.18	0.14	0.09	0.16	0.16	0.12	0.22	0.20	0.16	0.12	0.18	0.12	0.15	0.16	0.15	0.17	0.16	0.13	0.23	**0.13**	0.18
qwen2.5-vl-3b	0.20	0.16	0.11	0.19	0.16	0.11	0.18	0.14	0.10	0.24	0.18	0.12	0.22	0.22	0.14	0.26	0.23	0.15	0.32	0.26	0.18	0.32	0.20	0.14	0.24	0.19	0.13	0.23	0.21	0.20
qwen2.5-vl-7b	**0.34**	0.22	0.20	0.48	0.24	0.23	0.42	0.25	0.21	0.30	0.18	0.16	0.34	0.21	0.18	0.36	0.29	0.26	0.44	0.32	0.28	0.29	0.32	0.26	0.37	0.25	0.22	**0.18**	0.20	0.19
qwen2.5-vl-32b	0.32	0.18	0.20	**0.60**	0.26	0.32	0.40	0.20	0.24	0.44	0.30	0.31	0.50	0.30	0.32	0.40	0.28	0.29	0.55	0.34	0.37	0.57	0.26	0.30	0.47	0.26	0.29	0.21	0.20	0.18
qwen2.5-vl-72b	0.33	**0.26**	**0.26**	0.58	**0.46**	**0.46**	**0.48**	**0.44**	**0.42**	**0.50**	**0.44**	**0.41**	**0.58**	**0.39**	**0.41**	**0.55**	**0.46**	**0.45**	**0.66**	**0.52**	**0.54**	**0.62**	**0.52**	**0.52**	**0.50**	**0.44**	**0.43**	0.19	0.19	0.20

Table 2: The comparative performance of different models in data drift scenarios.



(a) Data drift of Qwen2.5-VL-7B under primary and secondary scenarios.

(b) Data drift of Qwen2.5-VL-32B under primary and secondary scenarios.

(c) Data drift of Qwen2.5-VL-72B under primary and secondary scenarios.

Figure 7: Visualization of data drift across different models. All models exhibit clear data drift, with the drift being more pronounced in secondary scenarios (right) than in primary ones (left).

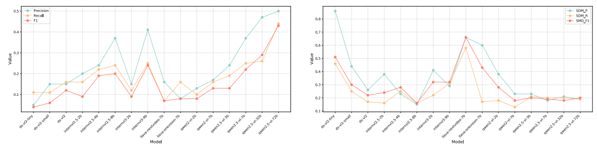
4.2 Experimental Findings

To evaluate the performance of current VLMs in short video advertising scenarios—particularly their ability to handle data drift and label drift—we conducted a series of detailed experiments and derived several key findings, which are elaborated in the following sections.

4.2.1 Data Drift

Conclusion 1: Current open-source VLMs perform moderately in short video advertising scenarios. We evaluated 16 mainstream open-source VLMs on data drift scenarios. Table 2 reports each model's monthly $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}_1$, and $\mathcal{SDM}$. The best-performing model was Qwen2.5-VL-72B, which, despite leading the group, only achieved $\overline{\mathcal{P}} = 0.50$, $\overline{\mathcal{R}} = 0.44$, and $\overline{\mathcal{F}_1} = 0.43$. Among models in the 7B–9B range, InternVL3-9B had the highest average performance with $\overline{\mathcal{P}} = 0.41$, $\overline{\mathcal{R}} = 0.25$, and $\overline{\mathcal{F}_1} = 0.24$.

A lower $\mathcal{SDM}$ indicates stronger robustness to data drift. InternVL2.5-8B, Qwen2-VL-7B, and the Qwen2.5-VL series showed strong stability, all with $\mathcal{SDM}_{\mathcal{F}_1}$ values below 0.2. Notably, although InternVL3-9B had the best average performance in the 7B–9B range, its $\mathcal{SDM}_{\mathcal{F}_1}$ was relatively high at 0.32, suggesting that strong risk identification ability does not necessarily imply strong robustness to data drift.



(a) Larger models exhibit stronger risk identification capabilities

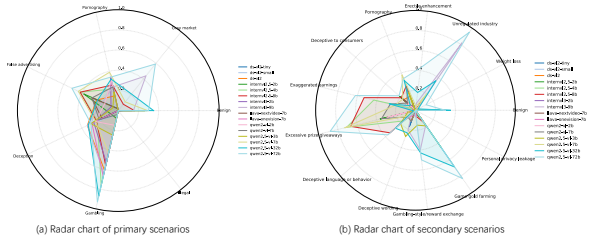
(b) Larger models demonstrate greater robustness to data drift

Figure 8: Models with larger parameter sizes exhibit stronger risk identification capabilities and greater robustness to data drift.

Conclusion 2: Current open-source VLMs exhibit limited robustness to data drift. Figure 7 illustrates the data drift trends across both primary and secondary risk categories on a monthly basis for the Qwen2.5-VL series and InternVL3-9B. Although larger models generally demonstrate stronger risk identification capabilities, none of them effectively mitigate the impact of data drift. This is reflected in substantial month-to-month variations in $\mathcal{P}$ and $\mathcal{R}$. For example, Qwen2.5-VL-72B shows a $\mathcal{P}$ gap as large as 0.33, increasing from 0.30 in 202408 to 0.63 in 202502.

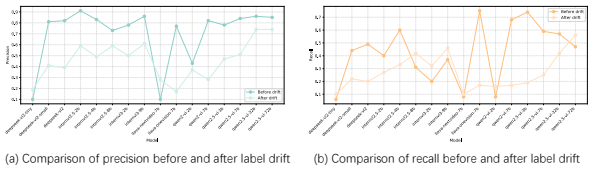
In Figure 7, lighter lines denote $\mathcal{P}$ and $\mathcal{R}$ across different primary and secondary categories, while darker lines represent the overall trend. The performance under secondary categories is notably weaker, partly due to the models' limited ability to recognize fine-grained risk scenarios. Additional visualizations of data drift patterns across different models are provided in Appendix B (Figure 14).

Conclusion 3: Models with larger parameter sizes demonstrate stronger capabilities in both risk scenario recognition and resistance to data drift. Figure 8(a) presents the $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ of different models across all scenarios and months. It reveals a clear positive correlation between model



(a) Radar chart of primary scenarios (b) Radar chart of secondary scenarios

Figure 9: Radar charts of data drift in primary and secondary scenarios.



(a) Comparison of precision before and after label drift (b) Comparison of recall before and after label drift

Figure 10: Visualization of model performance before and after label drift. Larger models demonstrate greater robustness to label drift.

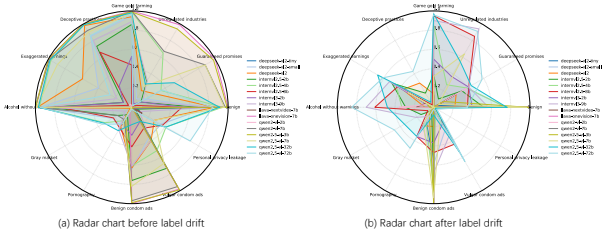
size and detection performance within the same model family. Figure 8(b) shows that models with larger parameter sizes tend to have lower $\mathcal{SDM}$, indicating better robustness to data drift. Notably, within the Qwen2.5-VL series, all variants exhibit relatively low $\mathcal{SDM}$, suggesting that this series as a whole is more resilient to data drift.

Conclusion 4: VLMs exhibit significant variability in risk identification performance across different scenarios. Figure 9 presents the $\mathcal{F}_1$ of various models under both primary and secondary categories. The results show substantial differences in model performance across scenarios. For example, Qwen2.5-VL-72B achieves an $\mathcal{F}_1$ of 0.93 in the gambling scenario, while its $\mathcal{F}_1$ drops to 0 in the illegal scenario. The performance gaps are even more pronounced in secondary scenarios.

Notably, all models fail to detect risks in the illegal category, which may be attributed to safety constraints imposed during the RLHF stage, where outputs related to illegal content are suppressed.

4.2.2 Label Drift

Conclusion 5: Label drift leads to performance degradation in nearly all models. Table 3 presents the $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ of different models before and after label drift across various scenarios. As shown, almost all models experience a performance drop following label drift. Taking Qwen2.5-VL-32B as an example, its $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}_1$ under lenient evaluation rules are 0.86, 0.57, and 0.58, respectively. After drift (under stricter auditing criteria), these metrics drop to 0.74, 0.42, and 0.44.



(a) Radar chart before label drift (b) Radar chart after label drift

Figure 11: Model performance under (a) lenient and (b) strict evaluation criteria

This reflects two key challenges: first, current VLMs struggle to identify fine-grained risk scenarios; second, their instruction-following capabilities still have room for improvement. Interestingly, Qwen2.5-VL-72B achieves a higher $\mathcal{F}_1$ after label drift than before. We interpret this as supporting evidence for Conclusion 3 that larger models possess stronger risk recognition capabilities. Additionally, the relatively low $\mathcal{F}_1$ of Qwen2.5-VL-72B before label drift suggests that its enhanced sensitivity to subtle risk cues may lead to lower scores under lenient evaluation settings, where such granularity is less rewarded.

Conclusion 6: Models with larger parameter sizes exhibit stronger robustness to label drift. Figure 10 presents the $\mathcal{P}$ and $\mathcal{R}$ of different models before and after label drift. As the number of parameters increases, the performance gap between the pre- and post-drift settings narrows significantly. Notably, the recall of Qwen2.5-VL-72B even achieves higher performance after label drift than before. As discussed in Conclusion 5, we believe this is primarily because large-parameter VLMs possess stronger capabilities in identifying fine-grained risk scenarios, which enables them to perform better under the stricter evaluation standards introduced by label drift.

Figure 11 shows radar charts of $\mathcal{F}_1$ before and after label drift. Following the drift, nearly all models experience significant drops in $\mathcal{F}_1$ across all scenarios, further demonstrating the adverse impact of label drift on models' risk identification capabilities. Detailed results are provided in Appendix C (Table 6 and 7).

4.2.3 Ablation Studies

We sampled 100 instances from each of the seven primary scenarios under data drift, resulting in a total of 700 examples, and conducted the following ablation experiments.

Conclusion 7: Incorporating ASR text di-

Model	$\overline{\mathcal{P}_\mathcal{L}}/\overline{\mathcal{P}_\mathcal{S}}$	$\overline{\mathcal{R}_\mathcal{L}}/\overline{\mathcal{R}_\mathcal{S}}$	$\overline{\mathcal{F}_\mathcal{L}}/\overline{\mathcal{F}_\mathcal{S}}$
deepseek-vl2-tiny	0.1 / 0.19	0.06 / 0.1	0.03 / 0.12
deepseek-vl2-small	0.81 / 0.41	0.44 / 0.22	0.5 / 0.24
deepseek-vl2	0.82 / 0.39	0.49 / 0.2	0.53 / 0.15
internvl2.5-2b	**0.91** / 0.59	0.4 / 0.27	0.45 / 0.27
internvl2.5-4b	0.83 / 0.49	0.6 / 0.33	0.66 / 0.34
internvl2.5-8b	0.73 / 0.59	0.31 / 0.42	0.33 / 0.41
internvl3-2b	0.78 / 0.5	0.2 / 0.32	0.25 / 0.26
internvl3-9b	0.86 / 0.61	0.37 / 0.46	0.46 / 0.45
llava-nextvideo-7b	0.1 / 0.28	0.08 / 0.1	0.03 / 0.08
llava-onevision-7b	0.77 / 0.17	**0.75** / 0.17	**0.75** / 0.17
qwen2-vl-2b	0.43 / 0.37	0.08 / 0.16	0.03 / 0.14
qwen2-vl-7b	0.82 / 0.28	0.68 / 0.17	0.72 / 0.18
qwen2.5-vl-3b	0.78 / 0.47	0.74 / 0.19	**0.75** / 0.21
qwen2.5-vl-7b	0.84 / 0.51	0.59 / 0.25	0.65 / 0.3
qwen2.5-vl-32b	0.86 / **0.74**	0.57 / 0.42	0.58 / 0.44
qwen2.5-vl-72b	0.85 / **0.74**	0.47 / **0.56**	0.46 / **0.53**

Table 3: Comparison of model performance under lenient and strict settings. $\overline{\mathcal{P}_\mathcal{L}}$ and $\overline{\mathcal{P}_\mathcal{S}}$ denote precision under lenient and strict evaluation criteria, respectively; the same applies to other metrics.

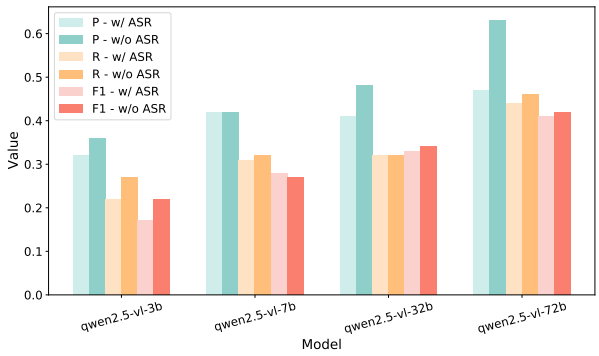


Figure 12: Visual comparison with and without ASR text. Directly inserting ASR text into the prompt tends to degrade model performance in most cases.

rectly into the prompt leads to degraded model performance. Intuitively, we expected that adding ASR text to the prompt would enhance VLMs' ability to identify risky content, essentially functioning as a form of multimodal fusion at the input level. However, the experimental results are counterintuitive. As illustrated in Figure 12, including ASR text noticeably harms model performance (see Appendix D (Table 8) for detailed results).

We attribute this surprising phenomenon to the fact that ASR text often occupies a large number of tokens, while the actual risk-related content typically consists of only a few tokens. The vast majority of the remaining tokens are irrelevant or benign. As a result, the ASR input introduces substantial token-level noise, making it more difficult for the model to accurately localize the few tokens that indicate violations.

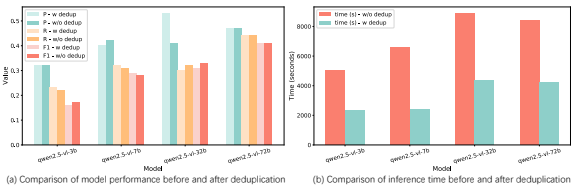


Figure 13: Comparison of model performance and inference time before and after deduplication.

Conclusion 8: Intra-video similarity-based deduplication does not degrade model performance but can significantly reduce inference time. Figure 13(a) provides a visual comparison of model performance before and after deduplication, showing minimal differences. In some cases deduplication even leads to improved performance.

We further analyzed the inference time of the models before and after deduplication, as illustrated in Figure 13(b) (see Appendix E (Table 9) for the corresponding comparison of frame counts). The results show a significant reduction in inference time after deduplication. This is primarily because, following deduplication, most videos contain fewer than 25 frames—the default number of input frames for the Qwen2.5-VL series.

5 Conclusion

This paper introduces ADB, the first benchmark specifically designed for short video advertising scenarios. We evaluate 16 open-source VLMs across two major types of distributional shifts—data drift and label drift. Our findings reveal that current open-source VLMs exhibit significant limitations in handling short video advertising content, particularly in their ability to cope with data and label drift. These shortcomings highlight two key challenges for existing VLMs: limited fine-grained semantic understanding and insufficient adherence to strict instruction following.

Limitations

Due to resource constraints, we did not evaluate the performance of commercial models such as GPT-4o. Additionally, this paper primarily focuses on identifying the problem: current open-source VLMs perform suboptimally under data drift and label drift scenarios. However, we do not propose specific solutions. In future work, we plan to build upon ADB and focus on improving the performance of VLMs in short video advertising tasks, particularly under data and label drift conditions.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. 2024b. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Bin Liu, Mingyu Wu, Minze Tao, Qin Wang, Luye He, Guoliang Shen, Kai Chen, and Junchi Yan. 2020. Video content analysis for compliance audit in finance and security industry. *Ieee Access*, 8:117888–117899.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llavanext: Improved reasoning, ocr, and world knowledge.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, YiFan Zhang, Shisong Tang, Changyi Liu, Haojie Ding, Kaiyu Jiang, and 1 others. 2025. Vlm as policy: Common-law content moderation framework for short video platform. *arXiv preprint arXiv:2504.14904*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online information review*, 43(1):72–88.

Dinh Duy Phan, Thanh Thien Nguyen, Quang Huy Nguyen, Hoang Loc Tran, Khac Ngoc Khoi Nguyen, and Duc Lung Vu. 2022. Lspd: A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15(1).

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich

social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Piotr Szwed, Pawel Skrzynski, and Wojciech Chmiel. 2016. Risk assessment for a video surveillance system based on fuzzy cognitive maps. *Multimedia Tools and Applications*, 75:10667–10690.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

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

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, and 1 others. 2024c. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, and 1 others. 2025. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Definition of Violation Scenarios

The detailed definitions of the 6 primary scenarios and 12 secondary scenarios for data drift are provided in Table 4. The definitions of the 10 scenarios for label drift are listed in Table 5.

B Detailed Visualization Results of Data Drift

Figure 14 shows the data drift patterns of 16 open-source VLMs across primary and secondary scenarios. While models with larger parameter sizes tend to exhibit stronger average risk identification capabilities, none of the models demonstrate satisfactory robustness to data drift.

C Detailed Experimental Results for Each Scenario under Label Drift

Tables 6 and 7 present detailed results of each model before and after label drift across different scenarios.

D Detailed Results of the Ablation Study

Table 8 presents the results of a comparative experiment on whether ASR text is included in the prompt. It shows that, in most cases, incorporating ASR text leads to a decline in model performance.

Table 9 presents the comparative results of whether inter-frame deduplication was applied. The results show that model performance differs little before and after deduplication. Table 10 illustrates the inference time before and after deduplication, demonstrating that deduplication significantly reduces inference time. All experiments were conducted on two H20 GPUs.

Figure 15 presents the distribution of video frames before and after deduplication. The average number of frames before deduplication is 42.1, while after deduplication it drops to 16.7, indicating that deduplication can significantly reduce inference cost.

E Detailed Configurations of Multimodal Small Models

Table 10 presents the detailed configurations of our trained multimodal small model, which is primarily used for preliminary data filtering to identify hard cases.

F Detailed Configurations of Open-Source VLMs

Table 11 presents the detailed configurations of the 16 open-source VLMs we evaluated. The input image resolution is $(364, 224)$, and all experiments were conducted on two H20 GPUs.

G Prompt Examples

Figure 16 shows the prompt under lenient rules, corresponding to the pre-label drift setting. Figure 17 presents the prompt under strict rules, corresponding to the post-label drift setting. Figures 19 and 18 display the prompts used in data drift scenarios.

Primary Scenarios	Secondary Scenarios	Defination
Deception	Deceptive language or behavior	[1] Misleading Language: Clickbait expressions such as "Totally shocked" or "Will be deleted if not watched now." [2] Misleading Interaction: Videos containing fake interactive elements, such as simulated incoming calls or fake pause buttons.
	Deceptive wording	[1] Guaranteed Claims: Any form of guarantee about product effectiveness, including those made in a personal capacity. [2] Hyped Sales Claims: Exaggerated expressions about sales volume, such as "best-seller" or "sold out instantly." [3] Fabricated Gimmicks to Induce Purchases: Phrases like "free treatment," "free gift," or "buy now, huge profit guaranteed."
False advertising	Deceptive to consumers	Exaggerated claims about product efficacy or functionality.
	Excessive prize giveaways	The value of the free gift exceeds that of the main product, or the gift's value is clearly exaggerated.
	Exaggerated earnings	Claims of earning large amounts of cash by playing games or watching videos, with statements such as "playing games or watching videos is more profitable than working a regular job."
Pornography	Pornography	[1] Revealing clothing with close-up shots of breasts, legs, or buttocks. [2] Text or language containing sexual innuendos. [3] Implicit depictions of sexual acts. [4] Animal sexual activity.
Gray market	Unregulated industry	Involves borderline sexually suggestive services such as sleep companionship, wake-up calls, paid gaming companionship, or paid chat interactions.
	Involves weight loss	Promotion of weight loss products, such as diet pills or slimming supplements.
	Involves erectile enhancement	Promotion of male enhancement products, such as aphrodisiacs or virility supplements.
Illegal	Personal privacy leakage	Disclosure of personal privacy information, such as ID numbers, license plate numbers, home addresses, and similar details.
Gambling	Gambling-style/reward exchange	Involves gambling-related content such as Mark Six lotteries, slot machines, and similar products.
	Game gold farming	Promotion of earning money by obtaining and selling in-game items through gameplay.

Table 4: Definitions of primary and secondary scenarios under data drift.

Scenarios	**Defination**
Guaranteed promises	Making guarantees about product effectiveness in a personal capacity or any form, with claims such as "guaranteed cure" or "guaranteed results."
Game gold farming	Promotion of earning money by obtaining and selling in-game items through gameplay.
Vulgar condom ads	Prolonged display of condom products in the video accompanied by sexually suggestive behavior. Mere display of external packaging without explicit or suggestive content is not considered a violation.
Alcohol without warnings	The video depicts alcohol consumption or features alcoholic products without displaying warning messages such as "Alcohol consumption is prohibited for minors."
Unregulated industries	Involves borderline sexually suggestive services such as sleep companionship, wake-up calls, paid gaming companionship, or paid chat interactions.
Exaggerated earnings	Claims of earning large amounts of cash by playing games or watching videos, with statements such as "playing games or watching videos is more profitable than working a regular job."
Deceptive practices	The video contains misleading interactive elements designed to trick users into clicking, such as fake pause buttons or simulated incoming call screens.
Personal privacy leakage	The content discloses personal privacy information such as ID numbers, home addresses, phone numbers, or license plate numbers.
Gray market	Promotion of products related to weight loss, breast enhancement, male enhancement, height increase, or body odor removal.
Pornography	[1] Revealing clothing with close-up shots of breasts, buttocks, or legs; [2] Sexually suggestive content in spoken language or written text; [3] Visuals that imply sexual acts; [4] Depiction of animal sexual activity.

Table 5: Definitions of scenarios under label drift.

	Alcohol without warnings			Gray market			Pornography			Benign condom ads			Vulgar condom ads			Personal privacy leakage		
	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$
deepseek-vl2-tiny	1.0 / 1.0	0.02 / 0.55	0.04 / 0.71	0.0 / 0.14	0.0 / 0.02	0.0 / 0.04	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
deepseek-vl2-small	1.0 / 1.0	0.66 / 0.62	0.79 / 0.76	1.0 / 0.27	0.09 / 0.69	0.17 / 0.39	1.0 / 0.0	0.05 / 0.0	0.1 / 0.0	1.0 / 0.75	0.9 / 0.24	0.95 / 0.36	1.0 / 0.0	0.74 / 0.0	0.85 / 0.0	1.0 / 0.3	0.25 / 0.13	0.4 / 0.18
deepseek-vl2	1.0 / 1.0	0.75 / 0.12	0.86 / 0.22	1.0 / 0.15	0.16 / 0.96	0.28 / 0.27	1.0 / 0.0	0.11 / 0.0	0.2 / 0.0	1.0 / 0.53	0.88 / 0.08	0.94 / 0.14	1.0 / 0.72	0.97 / 0.18	0.98 / 0.29	1.0 / 0.28	0.42 / 0.56	0.59 / 0.37
internvl2.5-2b	1.0 / 1.0	0.52 / 0.35	0.68 / 0.52	1.0 / 0.21	0.06 / 0.83	0.11 / 0.34	1.0 / 1.0	0.18 / 0.05	0.31 / 0.1	1.0 / 0.84	0.75 / 0.21	0.86 / 0.34	1.0 / 0.27	0.58 / 0.12	0.73 / 0.17	1.0 / 0.38	0.05 / 0.75	0.1 / 0.5
internvl2.5-4b	1.0 / 1.0	0.89 / 0.61	0.94 / 0.75	1.0 / 0.28	0.21 / 0.71	0.35 / 0.4	1.0 / 0.84	0.48 / 0.16	0.65 / 0.27	1.0 / 0.77	0.99 / 0.96	0.99 / 0.85	1.0 / 0.0	0.98 / 0.0	0.99 / 0.0	1.0 / 0.5	0.94 / 0.41	0.97 / 0.45
internvl2.5-8b	1.0 / 1.0	0.35 / 0.21	0.52 / 0.35	1.0 / 0.29	0.02 / 0.72	0.04 / 0.42	0.0 / 0.89	0.0 / 0.81	0.0 / 0.85	1.0 / 0.96	0.28 / 0.94	0.44 / 0.95	1.0 / 0.04	0.49 / 0.02	0.66 / 0.03	1.0 / 0.28	0.02 / 0.96	0.04 / 0.43
internvl3-2b	1.0 / 1.0	0.36 / 0.25	0.53 / 0.4	1.0 / 0.3	0.06 / 0.7	0.11 / 0.42	1.0 / 0.23	0.02 / 0.87	0.04 / 0.37	1.0 / 0.97	0.36 / 0.34	0.53 / 0.5	1.0 / 0.0	0.15 / 0.0	0.26 / 0.0	1.0 / 0.33	0.05 / 0.98	0.1 / 0.49
internvl3-9b	1.0 / 1.0	0.69 / 0.39	0.81 / 0.57	1.0 / 0.27	0.07 / 0.9	0.13 / 0.42	1.0 / 0.96	0.19 / 0.92	0.32 / 0.94	1.0 / 0.83	0.8 / 0.98	0.89 / 0.9	1.0 / 0.0	0.67 / 0.0	0.8 / 0.0	1.0 / 0.37	0.1 / 0.91	0.18 / 0.53
llava-nextvideo-7b	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.18	0.0 / 0.49	0.0 / 0.27	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 1.0	0.0 / 0.07	0.0 / 0.13	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.38	0.0 / 0.05	0.0 / 0.09
llava-onevision-7b	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	0.99 / 0.0	0.99 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0
qwen2-vl-2b	1.0 / 1.0	0.05 / 0.2	0.1 / 0.34	0.0 / 0.29	0.0 / 0.56	0.0 / 0.39	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	1.0 / 0.7	0.01 / 0.21	0.02 / 0.32	1.0 / 0.0	0.02 / 0.0	0.04 / 0.0	1.0 / 0.43	0.01 / 0.03	0.02 / 0.06
qwen2-vl-7b	1.0 / 1.0	0.95 / 0.98	0.97 / 0.99	1.0 / 0.0	0.78 / 0.0	0.88 / 0.0	1.0 / 0.0	0.5 / 0.0	0.67 / 0.0	1.0 / 0.37	1.0 / 0.07	1.0 / 0.12	1.0 / 0.0	0.85 / 0.0	0.92 / 0.0	1.0 / 1.0	0.99 / 0.03	0.99 / 0.06
qwen2.5-vl-3b	1.0 / 1.0	0.98 / 0.97	0.99 / 0.98	1.0 / 0.35	0.97 / 0.07	0.98 / 0.12	1.0 / 0.0	0.89 / 0.0	0.94 / 0.0	1.0 / 0.64	0.99 / 0.09	0.99 / 0.16	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	1.0 / 0.83	1.0 / 0.05	1.0 / 0.09
qwen2.5-vl-7b	1.0 / 1.0	0.92 / 0.93	0.96 / 0.96	1.0 / 0.75	0.78 / 0.03	0.88 / 0.06	1.0 / 0.96	0.14 / 0.48	0.25 / 0.64	1.0 / 0.51	0.97 / 0.23	0.98 / 0.32	1.0 / 0.0	0.82 / 0.0	0.9 / 0.0	1.0 / 0.41	0.8 / 0.24	0.89 / 0.3
qwen2.5-vl-32b	1.0 / 1.0	0.84 / 0.62	0.91 / 0.76	1.0 / 0.74	0.34 / 0.32	0.51 / 0.45	1.0 / 1.0	0.16 / 0.38	0.28 / 0.55	1.0 / 0.95	0.97 / 0.99	0.98 / 0.97	1.0 / 0.5	0.99 / 0.03	0.99 / 0.06	1.0 / 0.51	0.97 / 0.95	0.98 / 0.67
qwen2.5-vl-72b	1.0 / 1.0	0.72 / 0.21	0.84 / 0.35	1.0 / 0.45	0.22 / 0.82	0.36 / 0.58	1.0 / 0.98	0.01 / 0.85	0.02 / 0.91	1.0 / 0.93	0.94 / 0.99	0.97 / 0.96	1.0 / 0.56	0.75 / 0.35	0.86 / 0.43	1.0 / 0.45	0.15 / 0.94	0.26 / 0.61

Table 6: Model performance under label drift scenarios (part1).

	Alcohol without warnings			Gray market			Pornography			Benign condom ads			Vulgar condom ads			Personal privacy leakage		
	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$	$\mathcal{P}_\mathcal{L}/\mathcal{P}_\mathcal{S}$	$\mathcal{R}_\mathcal{L}/\mathcal{R}_\mathcal{S}$	$\mathcal{F}_\mathcal{L}/\mathcal{F}_\mathcal{S}$
deepseek-vl2-tiny	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.09 / 0.11	0.31 / 0.31	0.15 / 0.16	0.0 / 1.0	0.0 / 0.38	0.0 / 0.55	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.07 / 0.0	0.35 / 0.0	0.12 / 0.0
deepseek-vl2-small	1.0 / 1.0	0.91 / 0.03	0.95 / 0.06	0.19 / 0.0	0.29 / 0.0	0.23 / 0.0	0.15 / 0.17	0.4 / 0.36	0.22 / 0.23	1.0 / 1.0	0.47 / 0.42	0.64 / 0.59	1.0 / 0.47	0.22 / 0.17	0.36 / 0.25	0.43 / 0.0	0.3 / 0.0	0.35 / 0.0
deepseek-vl2	1.0 / 0.83	0.97 / 0.05	0.98 / 0.09	0.5 / 0.0	0.01 / 0.0	0.02 / 0.0	0.15 / 0.16	0.52 / 0.35	0.23 / 0.22	1.0 / 1.0	0.46 / 0.11	0.63 / 0.2	1.0 / 0.0	0.25 / 0.0	0.4 / 0.0	0.2 / 0.0	0.39 / 0.0	0.27 / 0.0
internvl2.5-2b	1.0 / 0.94	0.77 / 0.17	0.87 / 0.29	0.1 / 0.89	0.63 / 0.08	0.17 / 0.15	0.83 / 0.5	0.05 / 0.02	0.09 / 0.04	1.0 / 1.0	0.61 / 0.65	0.76 / 0.79	1.0 / 0.0	0.58 / 0.0	0.73 / 0.0	1.0 / 0.0	0.02 / 0.0	0.04 / 0.0
internvl2.5-4b	1.0 / 0.95	0.93 / 0.2	0.96 / 0.33	0.22 / 0.0	0.56 / 0.0	0.32 / 0.0	0.15 / 0.22	0.18 / 0.12	0.16 / 0.16	1.0 / 1.0	0.44 / 0.71	0.61 / 0.83	1.0 / 0.25	0.35 / 0.02	0.52 / 0.04	0.58 / 0.02	0.31 / 0.01	0.41 / 0.01
internvl2.5-8b	1.0 / 1.0	0.6 / 0.44	0.75 / 0.61	0.12 / 0.5	0.81 / 0.04	0.22 / 0.07	0.18 / 0.53	0.42 / 0.26	0.25 / 0.35	1.0 / 1.0	0.26 / 0.31	0.41 / 0.47	1.0 / 0.55	0.14 / 0.36	0.25 / 0.44	0.49 / 0.0	0.3 / 0.0	0.37 / 0.0
internvl3-2b	1.0 / 1.0	0.65 / 0.01	0.79 / 0.02	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.12 / 0.43	0.49 / 0.3	0.2 / 0.35	1.0 / 1.0	0.17 / 0.23	0.29 / 0.37	1.0 / 0.69	0.09 / 0.11	0.17 / 0.19	0.25 / 0.0	0.01 / 0.0	0.02 / 0.0
internvl3-9b	1.0 / 0.98	0.86 / 0.54	0.92 / 0.7	0.21 / 0.45	0.05 / 0.14	0.08 / 0.21	0.25 / 0.75	0.29 / 0.06	0.27 / 0.11	1.0 / 1.0	0.44 / 0.51	0.61 / 0.68	1.0 / 0.75	0.25 / 0.21	0.4 / 0.33	0.83 / 0.0	0.05 / 0.0	0.09 / 0.0
llava-nextvideo-7b	0.0 / 0.12	0.0 / 0.44	0.0 / 0.18	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.09 / 0.0	0.82 / 0.0	0.17 / 0.0	1.0 / 1.0	0.01 / 0.01	0.02 / 0.02	0.0 / 0.67	0.0 / 0.16	0.0 / 0.26	0.11 / 0.0	0.13 / 0.0	0.12 / 0.0
llava-onevision-7b	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.29 / 0.0	0.02 / 0.0	0.04 / 0.0	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0	1.0 / 0.0	0.97 / 0.0	0.98 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
qwen2-vl-2b	0.0 / 0.96	0.0 / 0.22	0.0 / 0.36	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.08 / 0.11	0.86 / 0.68	0.15 / 0.19	1.0 / 1.0	0.01 / 0.01	0.02 / 0.02	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.03 / 0.0	0.01 / 0.0	0.01 / 0.0
qwen2-vl-7b	1.0 / 0.0	1.0 / 0.0	1.0 / 0.0	0.25 / 0.0	0.13 / 0.0	0.17 / 0.0	0.14 / 0.0	0.15 / 0.0	0.15 / 0.0	1.0 / 1.0	0.95 / 1.0	0.97 / 1.0	1.0 / 0.0	0.91 / 0.0	0.95 / 0.0	0.5 / 0.0	0.01 / 0.0	0.02 / 0.0
qwen2.5-vl-3b	1.0 / 1.0	1.0 / 0.02	1.0 / 0.04	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.33 / 0.83	0.08 / 0.05	0.13 / 0.09	1.0 / 1.0	0.99 / 0.98	0.99 / 0.99	1.0 / 0.0	0.98 / 0.0	0.99 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
qwen2.5-vl-7b	1.0 / 0.0	0.98 / 0.0	0.99 / 0.0	0.0 / 0.67	0.0 / 0.02	0.0 / 0.04	0.13 / 0.33	0.28 / 0.16	0.18 / 0.21	1.0 / 1.0	0.69 / 0.88	0.82 / 0.94	1.0 / 0.44	0.5 / 0.08	0.67 / 0.14	0.94 / 0.0	0.16 / 0.0	0.27 / 0.0
qwen2.5-vl-32b	1.0 / 0.97	0.95 / 0.32	0.97 / 0.48	0.21 / 0.28	0.76 / 0.48	0.33 / 0.35	0.23 / 0.41	0.35 / 0.31	0.28 / 0.35	1.0 / 1.0	0.1 / 0.04	0.18 / 0.08	1.0 / 0.46	0.09 / 0.64	0.17 / 0.53	0.82 / 1.0	0.27 / 0.01	0.41 / 0.02
qwen2.5-vl-72b	1.0 / 0.97	0.91 / 0.76	0.95 / 0.85	0.22 / 0.45	0.86 / 0.49	0.35 / 0.47	0.13 / 0.58	0.43 / 0.42	0.2 / 0.49	1.0 / 1.0	0.01 / 0.01	0.02 / 0.02	1.0 / 0.51	0.02 / 0.9	0.04 / 0.65	0.89 / 1.0	0.57 / 0.02	0.7 / 0.04

Table 7: Model performance under label drift scenarios (part2).

Model	Benign			Gray market			Pornography			False advertising			Deception			Gambling			Illegal			Average		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$
qwen2.5-vl-3b (w/o asr)	0.23	0.5	0.31	0	0	0	0.59	0.16	0.25	0.5	0.03	0.06	0.19	0.76	0.31	1	0.47	0.64	0	0	0	0.36	0.27	0.22
qwen2.5-vl-3b (w asr)	0.23	0.5	0.31	0	0	0	0.57	0.12	0.2	0.29	0.05	0.09	0.17	0.7	0.27	1	0.17	0.3	0	0	0	0.32	0.22	0.17
qwen2.5-vl-7b (w/o asr)	0.23	0.81	0.36	0.9	0.09	0.16	0.54	0.38	0.44	0.25	0.52	0.34	0.2	0.04	0.07	0.85	0.4	0.54	0	0	0	0.42	0.32	0.27
qwen2.5-vl-7b (w asr)	0.27	0.57	0.36	0.88	0.07	0.13	0.55	0.34	0.42	0.25	0.68	0.37	0.19	0.21	0.2	0.82	0.31	0.45	0	0	0	0.42	0.31	0.28
qwen2.5-vl-32b (w/o asr)	0.34	0.4	0.37	0.6	0.18	0.28	0.55	0.26	0.35	0.23	0.17	0.2	0.23	0.47	0.31	0.93	0.78	0.85	0.5	0.01	0.02	0.48	0.32	0.34
qwen2.5-vl-32b (w asr)	0.42	0.36	0.39	0.43	0.13	0.2	0.59	0.23	0.33	0.28	0.13	0.18	0.24	0.58	0.34	0.91	0.82	0.86	0	0	0	0.41	0.32	0.33
qwen2.5-vl-72b (w/o asr)	0.37	0.4	0.39	0.45	0.55	0.5	0.77	0.23	0.35	0.29	0.88	0.43	0.54	0.21	0.3	0.99	0.94	0.96	1	0.01	0.02	0.63	0.46	0.42
qwen2.5-vl-72b (w asr)	0.35	0.3	0.32	0.54	0.58	0.56	0.86	0.24	0.38	0.26	0.84	0.4	0.32	0.24	0.27	0.99	0.86	0.92	0	0	0	0.47	0.44	0.41

Table 8: Comparison of performance with and without ASR text in the prompt.

Model	Benign			Gray market			Pornography			False advertising			Deception			Gambling			Illegal			Average		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$
qwen2.5-vl-3b (w/o dedup)	0.23	0.5	0.31	0	0	0	0.57	0.12	0.2	0.29	0.05	0.09	0.17	0.7	0.27	1	0.17	0.3	0	0	0	0.32	0.22	0.17
qwen2.5-vl-3b (w dedup)	0.22	0.61	0.33	0	0	0	0.67	0.08	0.14	0.23	0.03	0.05	0.19	0.71	0.29	0.94	0.17	0.29	0	0	0	0.32	0.23	0.16
qwen2.5-vl-7b (w/o dedup)	0.27	0.57	0.36	0.88	0.07	0.13	0.55	0.34	0.42	0.25	0.68	0.37	0.19	0.21	0.2	0.82	0.31	0.45	0	0	0	0.42	0.31	0.28
qwen2.5-vl-7b (w dedup)	0.26	0.66	0.37	0.76	0.22	0.34	0.6	0.29	0.39	0.25	0.64	0.36	0.27	0.2	0.23	0.69	0.24	0.36	0	0	0	0.4	0.32	0.29
qwen2.5-vl-32b (w/o dedup)	0.42	0.36	0.39	0.43	0.13	0.2	0.59	0.23	0.33	0.28	0.13	0.18	0.24	0.58	0.34	0.91	0.82	0.86	0	0	0	0.41	0.32	0.33
qwen2.5-vl-32b (w dedup)	0.43	0.37	0.4	0.48	0.15	0.23	0.48	0.14	0.22	0.19	0.1	0.13	0.22	0.54	0.31	0.91	0.81	0.86	1	0.01	0.02	0.53	0.3	0.31
qwen2.5-vl-72b (w/o dedup)	0.35	0.3	0.32	0.54	0.58	0.56	0.86	0.24	0.38	0.26	0.84	0.4	0.32	0.24	0.27	0.99	0.86	0.92	0	0	0	0.47	0.44	0.41
qwen2.5-vl-72b (w dedup)	0.35	0.31	0.33	0.51	0.55	0.53	0.79	0.26	0.39	0.28	0.86	0.42	0.36	0.25	0.29	0.99	0.87	0.93	0	0	0	0.47	0.44	0.41

Table 9: Comparison of performance with and without inter-frame deduplication.

Category	Hyperparameters
Batch Size	Training: 256, Testing: 256
Learning Rate	1×10^{-4}
Optimizer	Adam
Dropout Rate	0.5
Number of Epochs	200
Max Length for BERT	1024
Pre-trained Text Model	BERT (‘bert-base-chinese‘)
Pre-trained Vision Model	CLIP (‘clip-vit-base-patch32‘)
Custom Model	MultimodalModel (MLP combining BERT and CLIP features)
BERT (‘bert-base-chinese‘)	102M
CLIP (‘clip-vit-base-patch32‘)	149M
MultimodalModel	461,319
- text_projector	196,864
- image_projector	131,328
- mlp	131,328
Total Parameters	251,461,319

Table 10: Detailed parameter configurations of multimodal small models.

Model	Release	Version	Input Frames
deepseek-vl2-tiny	2024-12	deepseek-ai/deepseek-vl2-tiny	8
deepseek-vl2-small	2024-12	deepseek-ai/deepseek-vl2-small	
deepseek-vl2	2024-12	deepseek-ai/deepseek-vl2	
internvl2.5-2b	2024-11	OpenGVLab/InternVL2-2B	7
internvl2.5-4b	2024-11	OpenGVLab/InternVL2-4B	
internvl2.5-8b	2024-11	OpenGVLab/InternVL2-8B	
internvl3-2b	2025-04	OpenGVLab/InternVL3-2B	7
internvl3-9b	2025-04	OpenGVLab/InternVL3-9B	
llava-nextvideo-7b	2024-06	llava-hf/LLaVA-NeXT-Video-7B-hf	7
llava-onevision-7b	2024-09	llava-hf/llava-onevision-qwen2-7b-ov-hf	16
qwen2-vl-2b	2024-08	Qwen/Qwen2-VL-2B-Instruct	25
qwen2-vl-7b	2024-08	Qwen/Qwen2-VL-7B-Instruct	
qwen2.5-vl-3b	2025-02	Qwen/Qwen2.5-VL-3B-Instruct	25
qwen2.5-vl-7b	2025-02	Qwen/Qwen2.5-VL-7B-Instruct	
qwen2.5-vl-32b	2025-02	Qwen/Qwen2.5-VL-32B-Instruct	
qwen2.5-vl-72b	2025-02	Qwen/Qwen2.5-VL-72B-Instruct	

Table 11: Configurations of the 16 open-source VLMs we evaluated.

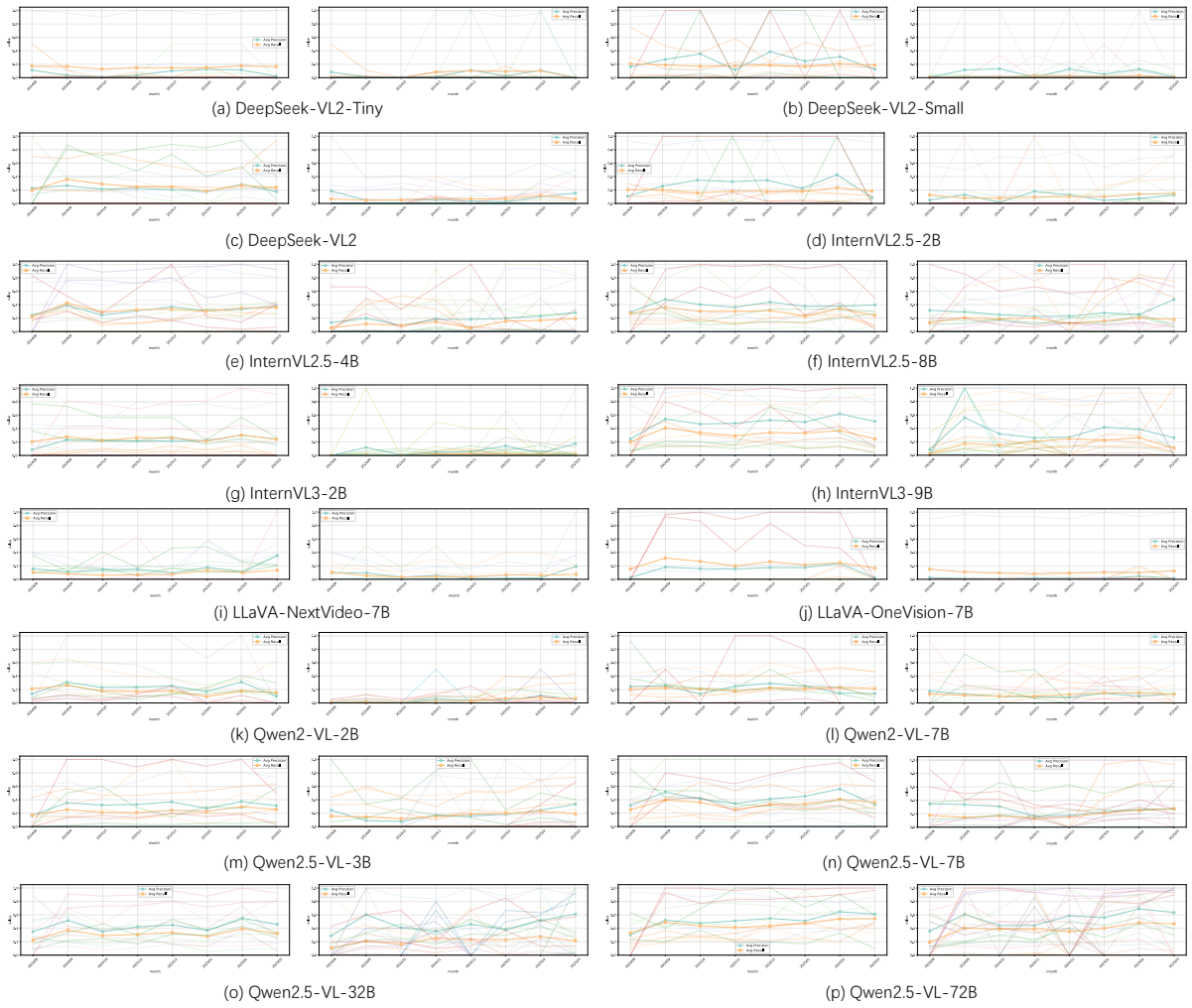


Figure 14: Data drift across different models: primary-level scenarios on the left, secondary-level scenarios on the right.

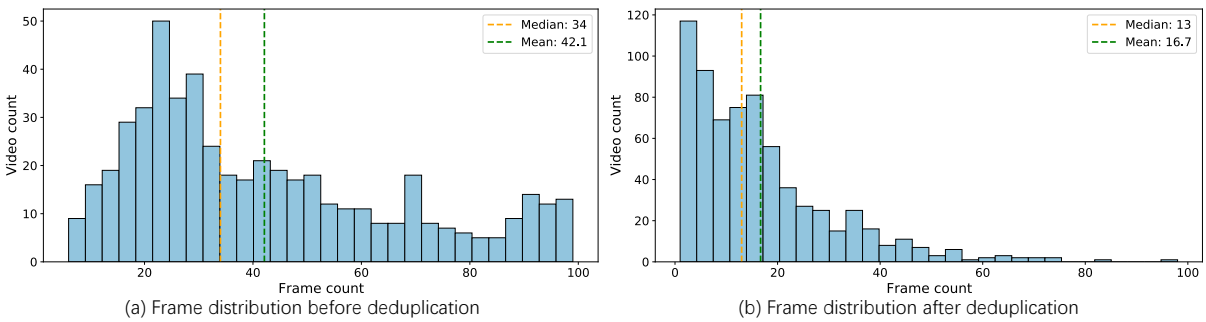


Figure 15: Frame distribution before and after inter-frame similarity-based deduplication.

请你扮演一名短视频广告审核员，职责是根据提供的视频帧判断视频是否存在以下违规行为，并从中选出唯一最合适的违规类型。具体违规类型及其定义如下：

违规类型及其定义
1. 非法-泄漏个人隐私：内容中泄露身份证号、家庭住址、电话号码、车牌号等个人隐私信息。
2. 灰产：宣传减肥、丰胸、壮阳、增高、除狐臭等相关产品。
3. 色情-低俗：
（1）穿着暴露，对胸、臀、腿等部位特写；
（2）语言或文字中含性暗示内容；
（3）暗示性行为的画面；
（4）展示动物性交内容。

如果视频不涉及上述任何问题，请判断为 “不违规” 。

输出形式
输出格式请使用以下 JSON 形式：
{"违规类型": "具体违规类型"}

(a) Chinese prompt under lenient label drift rules

Please act as a short video advertising content reviewer. Your task is to determine whether the provided video frames contain any of the following violations and identify the single most appropriate violation type. The specific types and their definitions are as follows:

Violation Types and Definitions
1. Personal privacy leakage: The content reveals personal information such as ID numbers, home addresses, phone numbers, or license plate numbers.
2. Gray market: Promotion of products related to weight loss, breast enhancement, male enhancement, height increase, or body odor removal.
3. Pornography:
(1) Revealing clothing with close-ups of breasts, buttocks, or legs;
(2) Sexually suggestive language or text;
(3) Visual implication of sexual acts;
(4) Depiction of animal sexual activity.

If the video does not contain any of the above issues, please classify it as 'Benign'.

Output Format
Please use the following JSON format for your output:
{ “Violation Type” : “Specific Violation Type” }

(a) English prompt under lenient label drift rules

Figure 16: Prompt example under lenient label drift rules.

请你扮演一名短视频广告审核员，职责是根据提供的视频帧判断视频是否存在以下违规行为，并从中选出唯一最合适的违规类型。具体违规类型及其定义如下：

违规类型及其定义
1. 涉及保证性承诺：以个人名义或任何形式担保产品效果，声称 “包治” 、 “保证见效” 等。
2. 涉及游戏打金：宣传打游戏能赚金币、爆装备、获得现金收益等。
3. 避孕套场景涉及低俗：视频中长时间展示避孕套本体，有色情暗示行为。若仅展示外包装且无过分内容，则不算违规。
4. 涉及饮酒且无警示语提示：视频中出现饮酒行为或酒类产品，但未出现 “未成年人禁止饮酒” 等警示语。
5. 涉及不规范行业：宣传哄睡、叫醒、付费陪玩/陪聊等陪伴类服务。
6. 涉及赚取金额夸大：宣传赠品价值高于商品本身，或鼓吹打游戏、看视频等轻松方式能轻松赚大钱。
7. 涉及误导操作：画面中出现误导用户点击的元素，如虚假暂停键、虚假来电界面等。
8. 非法-泄漏个人隐私：内容中泄露身份证号、家庭住址、电话号码、车牌号等个人隐私信息。
9. 灰产：宣传减肥、丰胸、壮阳、增高、除狐臭等相关产品。
10. 色情-低俗：
（1）穿着暴露，对胸、臀、腿等部位特写；
（2）语言或文字中含性暗示内容；
（3）暗示性行为的画面；
（4）展示动物性交内容。

如果视频不涉及上述任何问题，请判断为 “不违规” 。

输出形式
输出格式请使用以下 JSON 形式：
{"违规类型": "具体违规类型"}

(a) Chinese prompt under strict label drift rules

Please act as a short video advertising content reviewer. Your task is to determine whether the provided video frames contain any of the following violations and identify the single most appropriate violation type. The specific violation types and their definitions are as follows:

Violation Types and Definitions
1.Guaranteed promises: Providing guarantees about product effectiveness in a personal capacity or any form, using phrases like “guaranteed cure” or “guaranteed results.”
2.Game gold farming: Promoting the idea that playing games can earn in-game currency, rare items, or cash profits.
3.Vulgar condom ads: Prolonged display of condom products with sexually suggestive behavior. Displaying only the external packaging without explicit content is not considered a violation.
4.Alcohol without warnings: Depicting alcohol consumption or alcoholic products in the video without showing warning messages such as “Alcohol consumption is prohibited for minors.”
5.Unregulated industries: Promoting services such as sleep companionship, wake-up calls, paid gaming or chat companionship.
6.Exaggerated earnings: Claiming that the value of promotional gifts exceeds that of the main product, or exaggerating easy profits from playing games, watching videos, etc.
7.Deceptive practices: Displaying deceptive UI elements such as fake pause buttons or simulated incoming call interfaces intended to mislead users.
8.Personal privacy leakage: Revealing personal information such as ID numbers, home addresses, phone numbers, or license plate numbers.
9.Gray market: Promoting products related to weight loss, breast enhancement, male enhancement, height increase, or body odor removal.
10.Pornography:
(1) Revealing clothing with close-ups of breasts, buttocks, or legs;
(2) Sexually suggestive language or text;
(3) Visual implication of sexual acts;
(4) Depiction of animal sexual activity.

If the video does not contain any of the above issues, please classify it as 'Benign'.

Output Format
Please use the following JSON format for your output:
{ “Violation Type” : “Specific Violation Type” }

(a) English prompt under strict label drift rules

Figure 17: Prompt example under strict label drift rules.

你是一名短视频广告审核员，职责是根据视频内容判断其是否存在以下违规场景，并进一步识别对应的具体违规类型。请从下列定义中判断并选出唯一最合适的违规场景与违规类型。

违规场景
1. 误导
2. 虚假宣传
3. 色情
4. 灰产
5. 非法
6. 赌博

违规类型
（1）误导-涉及误导用语或行为：[1]误导用语：标题党描述，如“惊呆了”，“再不看就删除了”。[2]误导操作：视频中包含虚假动画按钮，如虚假手机来电、虚假暂停。
（2）误导-涉及诱导性文案：[1]涉及保证性承诺，以个人名义等任何形式担保产品效果。[2]炒作销量：涉及销量夸大用词，比如“卖爆了”等。[3]伪造噱头诱导用户购买：免费治疗、免费赠送、买到就是赚到。
（3）虚假宣传-欺骗消费者：宣传产品功效和功能涉及夸大描述
（4）虚假宣传-赠送奖品金额过高：赠品价值高于产品价值，或明显夸大赠品金额
（5）虚假宣传-赚取金额夸大：宣传靠打游戏、看视频赚了大量现金，踏实工作不如打游戏/看视频等描述。
（6）色情-低俗：[1]穿着暴露，对胸部、腿部、臀部进行特写。[2]文字、语言包含性暗示。[3]暗示性行为画面。[4]动物性交。
（7）灰产-不规范行业：涉及哄睡、叫醒、付费陪玩、付费陪聊等擦边业务
（8）灰产-涉及减肥：涉及减肥药等减肥相关产品
（9）灰产-涉及壮阳：设计壮阳药等壮阳产品
（10）非法-泄漏个人隐私：泄露个人隐私信息，如身份证好、车牌号、家庭地址等等
（11）赌博-棋牌捕鱼涉及赌博样式/实物兑换：涉及六合彩、老虎机等赌博相关产品
（12）赌博-游戏涉及打金：宣传通过打游戏爆装备赚钱

输出形式
最后给出的是json格式文件，包含“违规场景”和“违规类型”例如：
{"违规场景": "误导", "违规类型": "误导-涉及诱导性文案"}
若判断不涉及以上违规场景，则输出json为：
{"违规场景": "无", "违规类型": "无"}

Figure 18: Prompt example for data drift (Chinese version).

You are a short video advertising reviewer. Your responsibility is to assess the video content and determine whether it involves any of the following violation scenarios, and further identify the most appropriate specific violation type. Please select only one violation scenario and the corresponding violation type based on the definitions below.

Violation Scenarios
1.Deception
2.False advertising
3.Pornography
4.Gray market
5.Illegal
6.Gambling

Violation Types
(1) Deceptive language or behavior:
[1] Misleading language, such as clickbait expressions like "Totally shocked" or "Watch now before it' s deleted";
[2] Misleading interaction, such as fake animated buttons in the video (e.g., simulated incoming calls, fake pause buttons).
(2) Deceptive wording:
[1] Guaranteed claims made personally or in any form, such as "guaranteed cure," "guaranteed results";
[2] Hyped sales, using exaggerated terms like "best-seller," "sold out instantly";
[3] Fabricated gimmicks to induce purchases, such as "free treatment," "free gift," "buy now, big profit guaranteed."
(3) Deceptive to consumers:
Exaggerated claims about product efficacy or functionality.
(4) Excessive prize giveaways:
Free gift value exceeds that of the product itself or is clearly exaggerated.
(5) Exaggerated earnings:
Claims of earning large amounts of cash by playing games or watching videos, with phrases like "playing games is more profitable than working."
(6) Pornography:
[1] Revealing clothing with close-ups of breasts, legs, or buttocks;
[2] Text or language containing sexual innuendo;
[3] Visual implication of sexual acts;
[4] Depiction of animal sexual activity.
(7) Unregulated industry:
Includes sleep companionship, wake-up calls, paid gaming/chat companionship, and similar services with sexually suggestive undertones.
(8) Involves weight loss:
Promotion of diet pills or other weight loss-related products.
(9) Involves erectile enhancement:
Promotion of aphrodisiacs or products claiming to enhance male sexual performance.
(10) Personal privacy leakage:
Disclosure of personal privacy information such as ID numbers, license plate numbers, home addresses, etc.
(11) Gambling-style/reward exchange:
Content involving gambling-related products such as Mark Six lotteries, slot machines, etc.
(12) Game gold farming:
Claims that players can earn money through loot drops or in-game rewards.

Output Format
The final output should be in JSON format and include both "Violation Scenario" and "Violation Type", for example:
{"Violation Scenario": "Deception", "Violation Type": "Deceptive wording"}

If none of the above violation scenarios apply, the output should be:
{"Violation Scenario": "Benign", "Violation Type": "Benign"}

Figure 19: Prompt example for data drift (English version).