Reasoning-Enhanced Domain-Adaptive Pretraining of Multimodal Large Language Models for Short Video Content Governance

Zixuan Wang*, **Yu Sun***, **Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Dong, Zhuolin Hao, Hongyu Xiong, Yang Song**
TikTok Inc.
{zixuan.wang1, yu.sun, hongwei.w, baoyu.jing, xindong, haozhuolin, hongyu.xiong}@tiktok.com

Abstract

Short video platforms are evolving rapidly, making the identification of inappropriate content increasingly critical. Existing approaches typically train separate and small classification models for each type of issue, which requires extensive human-labeled data and lacks cross-issue generalization. We propose a reasoning-enhanced multimodal large language model (MLLM) pretraining paradigm for unified inappropriate content detection. To address the distribution gap between short video content and the original pretraining data of MLLMs, as well as the complex issue definitions, we introduce three targeted pretraining tasks: (1) *Caption*, to enhance the MLLM's perception of video details; (2) *Visual Question Answering (VQA)*, to deepen the MLLM's understanding of issue definitions and annotation guidelines; (3) *Chain-of-Thought (CoT)*, to enhance the MLLM's reasoning capability. Experimental results show that our pretraining approach significantly improves the MLLM's performance in both zero-shot and supervised fine-tuning (SFT) settings. In addition, our pretrained model demonstrates strong generalization capabilities to emergent, previously unseen issues.

1 Introduction

Short video platforms such as Reels and YouTube Shorts have experienced explosive growth. While these platforms offer unprecedented opportunities for communication and information dissemination, they have simultaneously facilitated the distribution of a wide range of inappropriate content. Videos featuring borderline sexual content, plagiarism, and fake engagement, not only affect user experience but also deteriorates content ecosystem in the long run. As a result, the regulation and governance of short video content have become increasingly critical to ensure a healthy digital ecosystem.

Current short video platforms typically implement a standardized content governance process, which consists of three sequential stages: The platform initially develops comprehensive *guidelines* for identifying specific *issues* (such as sexually suggestive content); Subsequently, human annotators are instructed to label video samples according to these guidelines; Finally, this labeled dataset serves as labeled data to fine-tune visual models (e.g., VLMo (Bao et al., 2022b), BEiT (Bao et al., 2022a), and X-VLM (Wang et al., 2022)) for video classification, which are then deployed for real-time content monitoring and filtering. However, this process comes with significant limitations: (1) *High manual labeling cost*. These visual models are typically small in size (usually only a few hundred megabytes) and lack world knowledge, so they require large volumes of manually labeled data, making the process time-consuming and expensive. (2) *No cross-issue generalization*. Each of these small models is designed to handle only a fixed issue and lacks the ability to generalize across different issues. Therefore, when the definition of an issue changes or a new issue emerges, the model must be retrained from scratch following the same labor-intensive process, leading to long development cycles.

To address the limitations of small specialized models, we explore the use of multi-modal large language models (MLLMs). MLLMs possess extensive world knowledge, which theoretically reduces the need for large-scale annotated data and offers better cross-issue generalization. However, applying MLLMs in this context presents two key challenges: (1) *Domain-specific data distribution*. The data on short video platforms exhibits unique characteristics, both in *format* (e.g., frame layouts, visual composition, and hashtag usage patterns) and in *semantics* (e.g., narrative styles, editing conventions, and internet slang). These characteristics often diverge significantly from the distribution of

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1104–1112

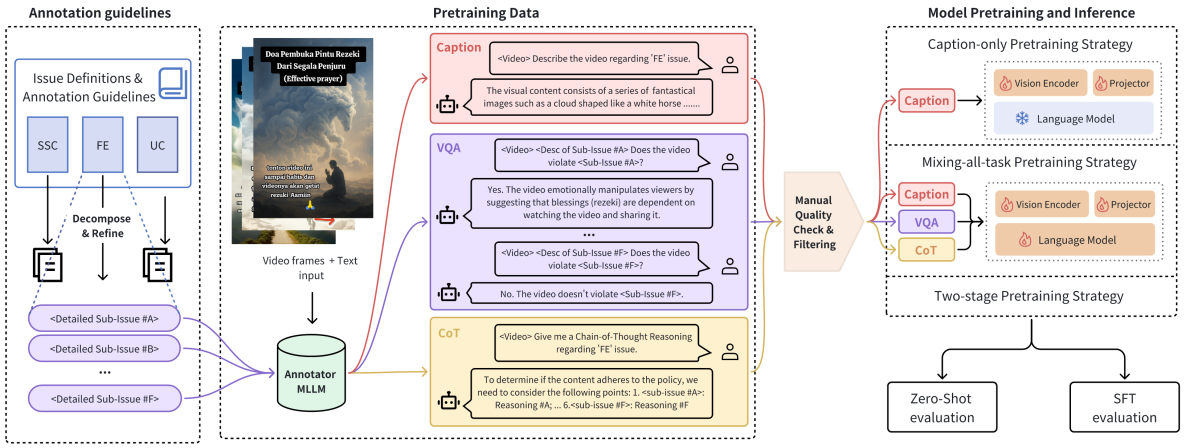

Figure 1: Illustration of our domain-adaptive pretraining approach for short video content governance. For each issue type, we first decompose the annotation guidelines into a set of sub-questions to assist pretraining data generation. An annotator MLLM is deployed to produce three types of pretraining data: Caption, VQA, and CoT, enabling the model to mimic human-like reasoning. The model can be pretrained using three different strategies, and the pretrained model is finally evaluated in both zero-shot and SFT settings.

data MLLMs were originally pretrained on, potentially impacting performance. (2) *Complex and evolving definition of issues*. Each content issue comes with its own well-defined criteria and decision rules, which are usually complicated and may evolve. For an MLLM to make accurate judgments, it must thoroughly understand and follow these guidelines to perform reliable inference.

To tackle the above challenges, we propose a novel domain-adaptive pretraining approach for short video governance. As illustrated in Figure 1, the pretraining approach includes three key components: (1) *Enhanced video detail perception*. To help MLLMs better adapt to the unique data patterns of short videos, we introduce a descriptive video captioning (*Caption*) task into the pretraining stage. These captions guide the model to attend to fine-grained video detail, especially those relevant to specific issues. (2) *Deep understanding of guidelines*. To enable MLLMs to thoroughly understand the nuances of issue annotation guidelines, we decompose each guideline into a set of sub-questions and construct a visual question answering (*VQA*) task. For example, in the case of borderline sexual content, the guideline is broken down into questions such as "`Are private body parts exposed`?", with annotated answers. This enables MLLMs to learn the specific judgment criteria for each issue. (3) *Structured reasoning*. Even after perceiving video details and understanding the guidelines, the model must follow a systematic reasoning process. To support this, we curate a chain-of-thought (*CoT*) task that teaches MLLMs to reason step by step following predefined logic flows. Together, the *Caption*, *VQA*, and *CoT* datasets provide end-to-end domain adaptation for MLLMs, which significantly improves the performance.

We conducted the pretraining on several open-source MLLMs, including LLaVA-OV (Li et al., 2024) and Qwen2.5-VL (Bai et al., 2024), using data from three content governance tasks: sexually suggestive content, unoriginal content, and fake engagement. Experimental results show that our domain-adaptive pretraining significantly improves model performance compared with native models, both in zero-shot and supervised fine-tuning (SFT) settings. Moreover, we demonstrate that our approach enables strong generalization to unseen issues with minimal supervision, highlighting its robustness and adaptivity. Our ablation studies further validate the effectiveness of our method across different model scales.

The contribution of this paper is as follows:

- We propose an MLLM pretraining framework for short video governance, demonstrating strong generalization to out-of-distribution issues not seen during pretraining.
- We design three pretraining tasks (Caption, VQA, CoT) to enhance the model's ability to perceive video details, understand annotation guidelines, and perform structured reasoning.
- Experiments show that our pretrained model

significantly improves performance in both zero-shot and SFT evaluation settings.

2 Related Work

2.1 Video Content Governance

Recent advances in deep neural networks have significantly enhanced content governance capabilities in detecting and classifying inappropriate video content on short video platforms (Yousaf and Nawaz, 2022; Li et al., 2025b). For example, multimodal features have been applied to identify misleading clickbait videos (Rahman et al., 2023; Sun et al., 2025; Liang et al., 2025); Age-adaptive learning techniques have been developed to detect and categorize inappropriate video content for different demographics of viewers, demonstrating sensitivity to visual needs (Alam et al., 2024; Zhang et al., 2024; Li et al., 2025a). These approaches underscore the potential of MLLMs in addressing the challenges of short video content governance.

2.2 Domain-Adaptive Pretraining

Recent research demonstrates the effectiveness of domain-adaptive pretraining in enhancing model performance (Song et al., 2025). Specifically, pretraining strategies have been used in model design, task formulation, and application scenarios (Guo et al., 2024). For example, in multimodal learning, efficient domain-specific pretraining has been applied to human activity recognition (Bulat et al., 2024) and short video understanding (Lu et al., 2025; Li et al., 2023); SimRAG (Hong et al., 2024) and domain-specific instruction tuning (Xie et al., 2024; Liu et al., 2024a; Wang et al., 2025; Dong et al., 2025) effectively align models with VQA tasks (Ging et al., 2024; Khullar et al., 2024); Adaptive techniques such as velocity-based domain reweighting (Luo et al., 2024; Leong et al., 2025), knowledge distillation (Shi et al., 2024) and structure-aware knowledge injection (Liu et al., 2024b) further improve knowledge retention and transfer.

3 The Proposed Pretraining Approach

3.1 A Unified Model for All Content Issues

Short video platforms often face a variety of content issues, such as non-original content. We denote the set of all issues of interest by $\mathcal{I}$. For each issue $i \in \mathcal{I}$, the policy team defines detailed descriptions and corresponding annotation guidelines, denoted by G_i. Human annotators are then trained using these guidelines to label a collection of videos, resulting in a dataset $\mathcal{D}_i$. This dataset is then used to fine-tune a visual classification model $\mathcal{M}_i$:

$$\mathcal{M}_i(v; G_i, \mathcal{D}_i) \rightarrow l, \quad \text{for } i \in \mathcal{I}, \tag{1}$$

where v is an input short video and l is the predicted label for v.

However, due to the lack of cross-issue generalization in these models, any change in the annotation guidelines G_i or the emergence of a new issue would require collecting a new human-annotated dataset $\mathcal{D}_i$ and retraining the model from scratch, which is extremely costly. Therefore, a promising solution is to leverage the broad, internalized knowledge of MLLMs to pretrain a unified model $\mathcal{M}$ that can generalize across different issues:

$$\mathcal{M}(v; \{G_i\}_{i \in \mathcal{I}}, \mathcal{D}) \rightarrow l, \tag{2}$$

where $\mathcal{D}$ is the pretraining dataset, which we will describe in detail in Section 3.3.

3.2 Annotation Guidelines Decomposition

Issue definitions and annotation guidelines are usually highly complex, typically covering dozens of scenarios, their exceptions, and both positive and negative illustrative examples. To help the model better understand these guidelines, it is essential to simplify and distill them into a set of sub-questions that are easier for the model to process. For instance, the annotation guidelines for the issue of sexually suggestive content consist of intricate definitions: "`Adult Image-Based Sexual Abuse occurs when the subject(s) depicted ...`". We decompose them into a collection of simple sub-questions such as "`Are private body parts exposed?`", "`Is there sexual teasing or invitation?`", and "`Are adult products shown?`". These sub-questions are later used to construct the pretraining dataset.

3.3 Pretraining Tasks

There are two main challenges in using MLLMs for short video content governance: (1) The unique characteristics of short videos limit the capabilities of native MLLMs to fully understand their content; (2) Issue annotation guidelines are often complex, making it difficult for native MLLMs to effectively follow and reason based on them. To address these challenges, we design three pretraining tasks:

Caption task is to help the model perceive fine-grained details of the input video, especially

those related to specific issues. As shown in Figure 1, the input of the caption task is a prompt such as "`Describe the video regarding the fake engagement issue.`", and the expected output is a 2–3 sentence description of the video from the perspective of fake engagement.

VQA task is to enable the model to develop a deep understanding of the annotation guidelines. To make more efficient use of the VQA data, we design two usage strategies: (1) *Binary QA*, where the input is a sub-question derived from the decomposed annotation guidelines described in Section 3.2. The output is a yes/no answer of whether the input video violates the given sub-question, as well as a detailed explanation. Examples of binary QA task are illustrated in Figure 1. (2) *Multi-choice QA*, where the input consists of all sub-questions, and the model is asked to select which issues the input video violates from the given options. It is important to note that VQA task relies on the model's accurate perception of video details obtained through the Caption task.

CoT task is to enable the model to integrate the answers from all sub-questions in the VQA task and produce a complete reasoning process leading to a final conclusion. As shown in Figure 1, for the fake engagement issue, the model should go through all the sub-questions. The final conclusion will be positive if any of the answer are positive.

3.4 Model Pretraining and Inference

How to organize the three pretraining tasks is also a critical question. A straightforward way is to mix all three tasks and jointly train the model. However, we found that the caption task is particularly important. Performing an initial round of caption-only training followed by joint training of all three tasks leads to better performance. We discuss the pretraining recipes in detail in Section 4.4.

We design two inference strategies. (1) The first is *zero-shot*, where the pretrained model directly classifies the input video and provides an explanation. We use the normalized logits of the predicted label token as the output probability. (2) The second strategy is *SFT*, where we add an MLP classification head on top of the model's final layer, and fine-tune the model using LoRA (Hu et al., 2022) on an additional SFT dataset for classification. It's worth noting that zero-shot inference offers more interpretable explanations for the model's predictions, while SFT inference is faster and currently serves as the solution for online deployment.

4 Experimental Setup

4.1 Content Issues

We collect the pretraining data and conduct experiments on real industrial data spanning three issues: Sexually Suggestive Content (**SSC**), Unoriginal Content (**UC**), and Fake Engagement (**FE**). To evaluate the model's ability to generalize to unseen or out-of-distribution issues, we also test the pretrained model on the Shocking Graphic Content (**SGC**) issue, which includes videos that cause physical discomfort to viewers. SGC issue is not included in the pretraining data.

4.2 Datasets

Pretraining data. We sample 50k positive and 50k negative videos for each of the 3 issues. An MLLM annotator is used to generate the Caption, VQA, and CoT tasks. The Caption task is generated independently, while the VQA and CoT tasks are generated simultaneously, resulting in a total of 920k instruction samples. After generation, we manually filter out any VQA-CoT samples that are inconsistent with the human annotations.

SFT data. We use human-annotated binary-labeled examples as SFT data for each issue to fine-tune the pretrained model, with 10% of the samples labeled as positive.

Evaluation data. We use 1k human-annotated samples as evaluation data for each issue, with 50% of the samples labeled as positive.

4.3 Baseline Models

We choose LLaVA-OV 7B (Li et al., 2024) and Qwen 2.5-VL 7B(Bai et al., 2024) as baseline MLLMs for continual pretraining. Both models consist of three components: a language model, a vision encoder, and a projector. Additionally, we compare our pretrained models with GPT-4o (Achiam et al., 2023) in the zero-shot setting.

4.4 Pretraining Strategy

We use three pretraining strategies:

Caption-only. In this strategy, the language model is frozen while the vision encoder and projector are trained using only the caption dataset. The goal is to enhance the model's ability to perceive fine-grained visual details.

Mixing-all-tasks. All three types of pretraining datasets are mixed and randomly shuffled. We train all components of the MLLM jointly on this mixed dataset.

Model	Pretraining strategy	SSC ACC	SSC F1	UC ACC	UC F1	FE ACC	FE F1	SGC ACC	SGC F1	Overall AUC
LLaVA-OV native	-	70.68	70.62	49.60	66.31	48.66	65.20	71.19	76.20	61.97
LLaVA-OV pretrained	Caption	74.53	75.38	65.67	71.82	48.09	64.95	71.19	75.31	70.89
	Mix	**82.13**	82.34	78.87	79.10	75.81	**77.39**	75.13	77.49	80.75
	Stage	80.65	81.99	**80.46**	**80.74**	71.13	72.69	**77.20**	**78.76**	**81.22**
Qwen2.5-VL native	-	74.23	76.12	51.69	67.12	53.82	66.44	74.92	78.04	69.49
Qwen2.5-VL pretrained	Caption	76.51	76.53	51.59	66.53	49.33	65.22	73.89	76.14	68.82
	Mix	81.64	**82.49**	75.40	75.59	73.33	74.33	72.12	76.17	74.10
	Stage	80.75	81.55	76.09	76.40	**76.00**	75.12	73.89	76.00	74.24
GPT-4o	-	75.71	72.97	65.97	67.24	64.15	68.97	74.62	78.09	-

Table 1: Result of zero-shot evaluation. Metrics are accuracy and F1 (in %). The last column is the overall ROC-AUC (in %) across all issues. The highest value in each column is shown in **bold**, and the second-highest is underlined. The overall AUC of GPT-4o cannot be computed because we cannot access its output probabilities.

Model	SSC AUC	SSC ACC	SSC P@R90	UC AUC	UC ACC	UC P@R90	FE AUC	FE ACC	FE P@R90	SGC AUC	SGC ACC	SGC P@R90
Native	92.97	84.05	79.76	80.98	71.53	59.92	91.38	84.35	75.70	99.18	89.86	97.95
Pretrained w/ Mix	**93.29**	**85.52**	**81.57**	**87.53**	**75.79**	**69.88**	**91.94**	**84.46**	**78.31**	99.26	94.93	**98.17**
Pretrained w/ Stage	93.00	84.84	80.57	86.65	73.41	64.84	91.88	84.34	76.12	**99.31**	**95.14**	98.05

Table 2: Result of SFT evaluation. Metrics are ROC-AUC, Accuracy, and P@R90 (Precision at 90% Recall) (in %). Following the actual model deployment setup, LLaVA-OV is used as the base model for SSC, FE, and SGC, while Qwen2.5-VL is used as the base model for UC. The highest value in each column is shown in **bold**.

Two-stage. This is a two-stage approach that combines the previous two strategies. We first apply Caption-only strategy to train the vision encoder and projector, and select the best checkpoint across epochs. Then, we switch to the Mixing-all-tasks strategy to fine-tune the entire model.

5 Experimental Results

5.1 Zero-Shot Evaluation

In-domain issues. The results of zero-shot evaluation are shown in Table 1. For the three in-domain issues: SSC, UC, and FE, the performance of the pretrained models consistently surpasses that of their native counterparts. Notably, the pretrained models also significantly outperform GPT-4o. This indicates that, despite GPT-4o's strong multimodal capabilities, it still struggles to fully understand complex issue annotation guidelines.

From an issue-specific perspective, the accuracy improvements brought by pretraining on SSC, UC, and FE are approximately 7–11%, 24%–31%, and 22%, respectively. The gains on UC and FE are greatly higher than those on SSC. This may be because SSC is detecting sexually suggestive content, a task for which native models already possess sufficient prior knowledge. In contrast, UC and FE are non-original content and fake engagement, whose precise definitions are given in the annotation guidelines. Native models lack strong priors for these issues. This further demonstrates the effectiveness of our VQA and CoT pretraining tasks in injecting domain-specific knowledge into the base models.

Out-of-domain issue. To demonstrate the generalization ability of the pretrained models, we also evaluate their performance on the SGC issue, which is not included in the pretraining data. The results show that pretraining significantly improves LLaVA's performance on this out-of-domain issue (6% absolute accuracy improvement).

Comparison between LLaVA and Qwen. On the three in-domain issues, pretraining brings accuracy improvements of 11%–31% for LLaVA and 7%–24% for Qwen. This suggests that the benefits of pretraining are more significant for LLaVA. Upon closer inspection, we found that one possible reason is that Qwen sometimes produces inconsistencies between its reasoning and final answers. That is, it may generate a correct and well-structured reasoning process, but still output an incorrect final prediction. We provide a detailed analysis example in the Appendix.

Pretraining strategy. Comparing the three pretraining strategies, we find that Two-stage consistently outperforms Caption-only and Mixing-all-

Configuration	AUC	ACC	P@R90
LLaVA-OV native + 100% SFT data	99.18	89.86	97.95
LLaVA-OV pretrained + 100% SFT data	**99.31**	95.14	98.05
LLaVA-OV pretrained + 50% SFT data	99.26	**95.35**	**98.17**

Table 3: ROC-AUC, Accuracy, and P@R90 scores (in %) on the SGC issue under SFT evaluation. The pretraining strategy is Stage. The highest value in each column is shown in **bold**.

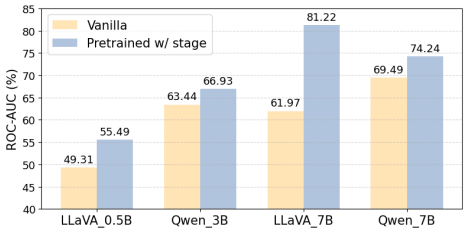

Figure 2: ROC-AUC of vanilla and pretrained models with different model sizes under zero-shot evaluation.

tasks in most cases. This suggests that first training the model to perceive fine-grained video details, followed by learning the annotation guidelines and reasoning process, is the most effective practice.

5.2 SFT Evaluation

The SFT evaluation results are presented in Table 2. Compared to zero-shot evaluation, most of our conclusions remain consistent, with a few differences:

(1) In most cases, pretraining still improves model performance under SFT evaluation, but the gains are generally smaller, typically around 1% to 10%. This is because the pretraining paradigm aligns closely with the zero-shot evaluation format (i.e., natural language generation), whereas the SFT evaluation is direct classification through probability outputs. This mismatch reduces the impact of pretraining under SFT evaluation.

(2) Regarding pretraining strategies, Mixing-all-tasks performs slightly better than Two-Stage. Note that Two-Stage includes an additional round of caption-only training. This suggests that adding an extra caption phase does not significantly enhance performance under SFT evaluation. The reason is similar to the above: the Caption task aligns more closely with natural language generation and contributes less to performance when the evaluation requires direct classification.

5.3 Ablation Study

Cross-issue generalization. Pretrained MLLMs exhibit strong cross-issue generalization capabilities. This generalization is reflected not only in the zero-shot setting, where pretrained models perform well on out-of-domain issues (as shown in the SGC issue in Table 1), but also in the SFT setting, where pretrained models can achieve or even surpass the performance of native models using significantly less SFT data. To demonstrate this, we fine-tune the pretrained LLaVA model using only 50% of the SFT data. The results, shown in Table 3, reveal that the pretrained LLaVA with just half the SFT data outperforms the native LLaVA trained with the full SFT dataset (compare row 3 with row 1). This highlights that the pretrained model possesses substantially stronger cross-issue generalization capabilities than the native model, enabling it to adapt to new issues with minimal SFT data.

Base model size. To investigate the impact of base model size on pretraining effectiveness, we additionally conduct pretraining and zero-shot evaluation on 0.5B versions of LLaVA and Qwen. The results in Figure 2 demonstrate that our pretraining approach also yields significant performance improvements (3%–6%) for 0.5B models. Moreover, we observe that the performance gains on 7B models (5%–19%) are higher than those on 0.5B models. This may be because larger models have greater capacity and can benefit more from the same amount of pretraining data, allowing them to reach a higher performance ceiling.

5.4 Case Study

We conducted a case study to intuitively demonstrate the effect of pretraining. As shown in Figure 3, the Spanish text in the video frame contains an obvious sexually suggestive message. However, the model without pretraining fails to recognize the meaning of the text and subsequently produces incorrect reasoning. In contrast, the pretrained model provides both an accurate answer and a correct reasoning process.

6 Conclusion

In this work, we propose a domain-adaptive pretraining approach for short video content governance. We design three novel pretraining tasks: Caption, VQA, and CoT, which effectively enhance MLLMs' ability to perceive video details, understand annotation guidelines, and perform reasoning. Experimental results show that our pretraining method significantly improves the model's performance on both zero-shot and SFT evaluations. Fur-

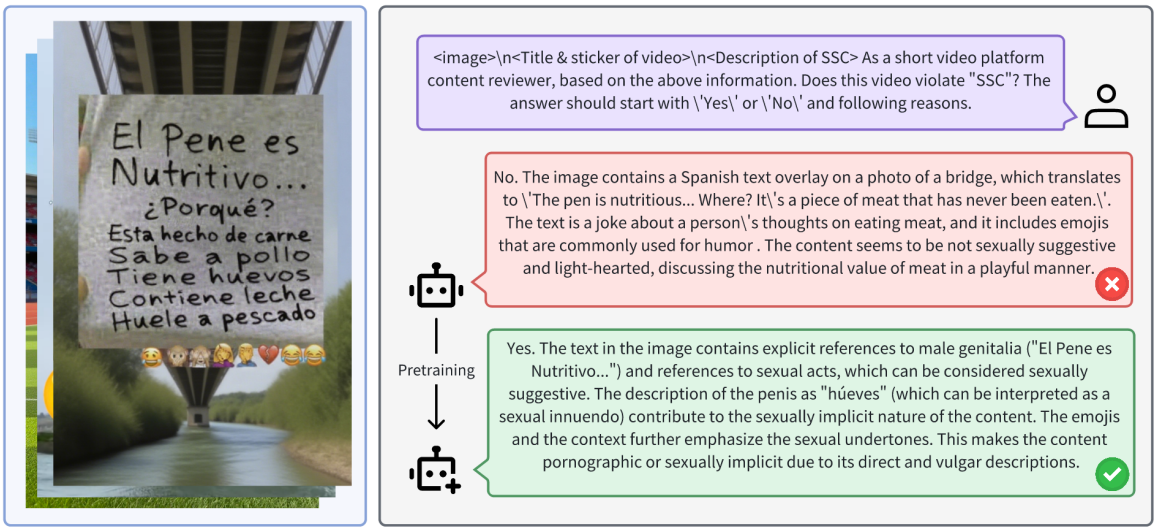

Figure 3: Left: An illustrative example of short video frames that violates the SSC policy. Right: Zero-shot evaluation prompt and the MLLM's responses before and after our pretraining.

thermore, the pretrained models exhibit strong generalization to new issues, which can substantially reduce manual annotation costs and shorten development cycles in real-world deployment scenarios.

7 Limitation

In this work, the largest model we used is 7B due to limitations in resources. Although the pretrained 7B models already perform well for short video content governance, we are still interested in exploring the impact of pretraining on even larger models. Regarding data types used for pretraining, in addition to MLLM-annotated data, we also plan to incorporate a large amount of unlabeled data to further enhance performance. Moreover, when decomposing annotation guidelines into sub-questions, we currently rely on manual design, which may not be optimal. In the future, we plan to explore reinforcement learning methods to automatically decompose annotation guidelines in a more efficient and adaptive manner.

References

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

Iftikhar Alam, Abdul Basit, and Riaz Ahmad Ziar. 2024. Utilizing age-adaptive deep learning approaches for detecting inappropriate video content. *Human Behavior and Emerging Technologies*, 2024(1):7004031.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Fei Huang, Qi Zhang, Yao Xiao, Dayiheng Liu, Chengyuan Li, Wendi Zheng, and Bo Zheng. 2024. Qwen2.5-VL technical report. *arXiv preprint arXiv:2405.17768*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022a. BEiT: BERT pre-training of image transformers. *Preprint*, arXiv:2106.08254.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022b. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Adrian Bulat, Yassine Ouali, Ricardo Guerrero, Brais Martinez, and Georgios Tzimiropoulos. 2024. Efficient vision-language pre-training via domain-specific learning for human activities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7978–8000.

Xin Dong, Sen Jia, Ming Rui Wang, Yan Li, Zhenheng Yang, Bingfeng Deng, and Hongyu Xiong. 2025. Coef-vq: Cost-efficient video quality understanding through a cascaded multimodal llm framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 4387–4395, New York, NY, USA. Association for Computing Machinery.

Simon Ging, María A Bravo, and Thomas Brox. 2024. Open-ended VQA benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. *arXiv preprint arXiv:2402.07270*.

Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao, and Yikang Shen. 2024. Efficient continual pre-training

by mitigating the stability gap. *arXiv preprint arXiv:2406.14833*.

Seongtae Hong, Joong Shin, Jaehyung Seo, Taemin Lee, Jeongbae Park, Cho Young, Byeongho Choi, and Heui-Seok Lim. 2024. Intelligent predictive maintenance rag framework for power plants: Enhancing qa with styledfs and domain specific instruction tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 805–820.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Dipika Khullar, Emmett Goodman, and Negin Sokhandan. 2024. Improved few-shot image classification through multiple-choice questions. *arXiv preprint arXiv:2407.16145*.

Hui Yi Leong, Yuheng Li, Yuqing Wu, Wenwen Ouyang, Wei Zhu, and Jiechao Gao. 2025. Amas: Adaptively determining communication topology for llm-based multi-agent system. *Preprint*, arXiv:2510.01617.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Wei Li, Bing Hu, Rui Shao, Leyang Shen, and Liqiang Nie. 2025a. Lion-fs: Fast & slow video-language thinker as online video assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3240–3251.

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. 2025b. Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification. *arXiv preprint arXiv:2508.21046*.

Yuyuan Li, Chaochao Chen, Yizhao Zhang, Weiming Liu, Lingjuan Lyu, Xiaolin Zheng, Dan Meng, and Jun Wang. 2023. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *Advances in Neural Information Processing Systems*, 36:12611–12625.

Hanzhong Liang, Jinghao Shi, Xiang Shen, Wang Zixuan, Wen Vera, Mehrani Ardalan, Chen Zhiqian, Wu Yifan, and Zhang Zhixin. 2025. Embedding-based retrieval in multimodal content moderation. *arXiv preprint arXiv:2507.01066*.

Gorden Liu, Yu Sun, Ruixiao Sun, Xin Dong, and Hongyu Xiong. 2024a. Agentps: Agentic process supervision for multi-modal content quality assurance through multi-round qa. *arXiv preprint arXiv:2412.15251*.

Kai Liu, Ze Chen, Zhihang Fu, Wei Zhang, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, and Jieping Ye. 2024b. Structure-aware domain knowledge injection for large language models. *arXiv preprint arXiv:2407.16724*.

Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, YiFan Zhang, Shisong Tang, Changyi Liu, Haojie Ding, Kaiyu Jiang, Kaiyu Tang, Bin Wen, Hai-Tao Zheng, Fan Yang, Tingting Gao, Di Zhang, and Kun Gai. 2025. VLM as policy: Common-law content moderation framework for short video platform. *Preprint*, arXiv:2504.14904.

Zheheng Luo, Xin Zhang, Xiao Liu, Haoling Li, Yeyun Gong, Chen Qi, and Peng Cheng. 2024. Velocitune: A velocity-based dynamic domain reweighting method for continual pre-training. *arXiv preprint arXiv:2411.14318*.

Sheikh Sowmen Rahman, Avishek Das, Omar Sharif, and Mohammed Moshiul Hoque. 2023. Identification of deceptive clickbait youtube videos using multimodal features. In *International Conference on Intelligent Computing & Optimization*, pages 199–208. Springer.

Jinghao Shi, Xiang Shen, Kaili Zhao, Xuedong Wang, Vera Wen, Zixuan Wang, Yifan Wu, and Zhixin Zhang. 2024. Cpfd: Confidence-aware privileged feature distillation for short video classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4866–4873, New York, NY, USA. Association for Computing Machinery.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025. Injecting domain-specific knowledge into large language models: a comprehensive survey. *arXiv preprint arXiv:2502.10708*.

Yu Sun, Yin Li, Ruixiao Sun, Chunhui Liu, Fangming Zhou, Ze Jin, Linjie Wang, Xiang Shen, Zhuolin Hao, and Hongyu Xiong. 2025. Audio-enhanced vision-language modeling with latent space broadening for high quality data expansion. *arXiv preprint arXiv:2503.17551*.

Wenguan Wang, Lin Ma, Shaodi You, Nianyi Jiang, Jianqiang Zhang, Zhedong Hu, Jianbing Shen, and Ling Shao. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2203.14962*.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. 2025. Filter-and-refine: A MLLM based cascade system for industrial-scale video content moderation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 873–880, Vienna, Austria. Association for Computational Linguistics.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the*

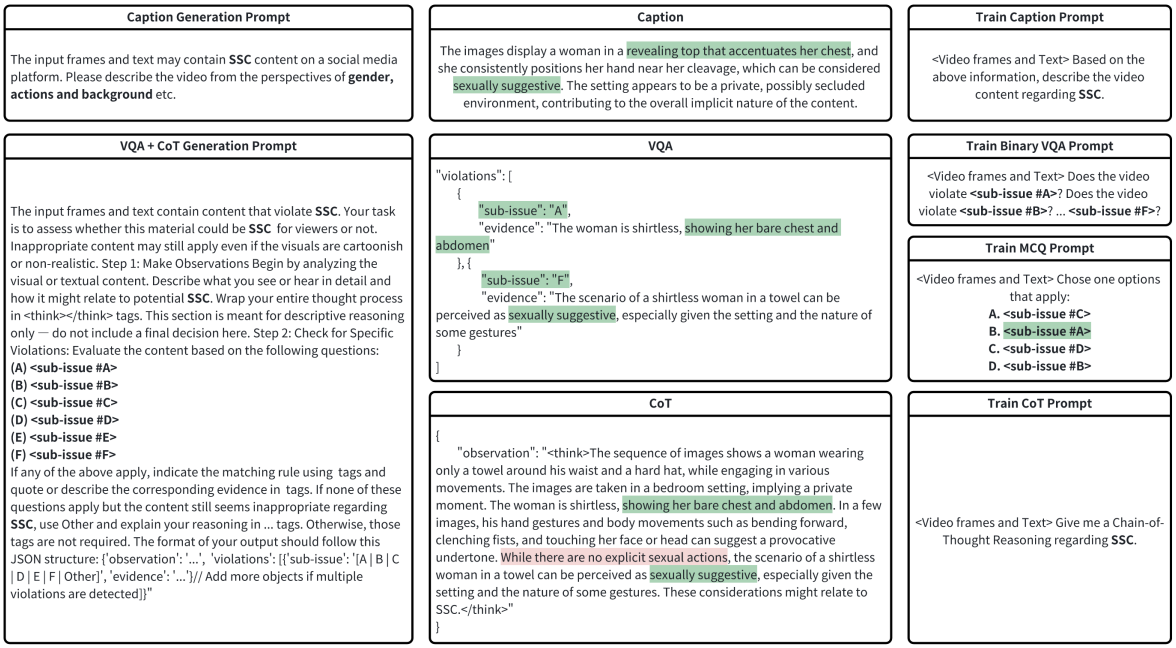

Figure 4: An illustrative example of prompts and the generated pretraining data. The first column is the prompt for generating the pretraining data. The second column is the output generated by the prompt in the first column. The third column is the prompt used during pretraining.

Association for Computational Linguistics ACL 2024, pages 10184–10201.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, 10:16283–16298.

Rongchao Zhang, Yiwei Lou, Dexuan Xu, Yongzhi Cao, Hanpin Wang, and Yu Huang. 2024. A learnable discrete-prior fusion autoencoder with contrastive learning for tabular data synthesis. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 16803–16811. AAAI Press.

A Details of Prompt Design

In Figure 4, we present a complete pipeline example for constructing pretraining data, including the prompts used for generating pretraining data, the generated outputs, and the prompts used during pretraining. To improve clarity, we highlight the issue and its related content using bold text. Additionally, we mark segments of the generated pretraining data that align with the sub-issue in green, and those that do not in red. This clearly illustrates that issue-specific knowledge is consistently embedded across all tasks.

B Analysis on Qwen's Inconsistency Issue

During zero-shot inference, we observed that when the answer format is designed as "Reason after Answer", Qwen often produces incorrect final answers despite providing correct reasoning. For example, when given a video that does not violate the UC policy and asked whether it violates the UC policy, the model responds: "Yes. The images appear to be original user-generated content without any visible watermarks indicating ownership by another entity..." Although the model clearly identifies the content as original, its final answer is still "Yes", and the probability score for the first token exceeds 0.67. As a result, evaluating performance based solely on the first token's probability leads to an underestimation of the model's true reasoning ability.