IPS: In-Prompt Process Supervision for Short Video Content Moderation

Mingchao Liu **Yu Sun** **Ruixiao Sun** **Xin Dong**
Xiang Shen **Hongwei Wang** **Hongyu Xiong** **Yang Song**
TikTok, Inc.
{gorden.liu, yu.sun, ruixiao.sun, xindong, xiang.shen, hongwei.w, hongyu.xiong}@tiktok.com
ys@sonyis.me

Abstract

Multimodal large language models (MLLMs) are effective at capturing the semantics of short video content; however, they often fail to attend to the policy-specific details required for reliable content moderation. To address this limitation, we introduce *IPS*, a novel framework that integrates In-prompt Process Supervision into MLLMs by introducing sequential reasoning over ancillary questions during fine-tuning. *IPS* consistently outperforms baseline MLLMs on public and proprietary benchmarks. Moreover, replacing human-annotated ancillary labels with MLLM-generated ones results in only marginal performance degradation, demonstrating robustness to noisy supervision and strong scalability with model-generated annotations. These findings establish *IPS* as a scalable and effective solution for complex multimodal classification in large-scale industrial settings.

1 Introduction

The rapid advancement of LLMs and MLLMs, such as GPT (Achiam et al., 2023), Gemini (Team et al., 2023), LLaVA (Li et al., 2024), and Qwen (Wang et al., 2024b; Team, 2025), has demonstrated remarkable capabilities across a wide range of applications, including visual question answering and contextual understanding.

Despite these advances, the effectiveness of MLLMs in highly specialized or sensitive domains remains underexplored. Standard end-to-end supervised fine-tuning (SFT) often falls short of meeting the nuanced requirements of domain-specific tasks. Unlike general content understanding tasks where MLLMs can leverage broad world knowledge, content moderation demands sophisticated reasoning grounded in complex and detail-intensive governance policies. The intricacy of these policies, combined with the challenges of consistent interpretation, presents substantial difficulties even for human annotators.

For instance, determining whether a social media post constitutes unoriginal content requires nuanced judgment. Human annotators follow predefined criteria by answering structured questions (e.g., whether the content contains copyrighted material or meaningful user-generated edits). In this way, a movie clip that would typically be labeled unoriginal may instead be considered original if the creator adds substantive commentary or humor. Such scenarios underscore the need for structured reasoning and fine-grained decision-making. (Lan et al., 2025)

In this paper, we propose a novel and effective approach to enhance the domain-specific supervised fine-tuning of MLLMs for complex content moderation tasks. Our work introduces two key innovations:

***IPS* framework**. We present *IPS*, a structured process supervision framework that aligns MLLM reasoning with human annotation workflows. *IPS* leverages ancillary labels to sequentially guide the model's decision-making. Specifically, each training instance embeds multiple question–answer pairs in sequence, where `<ans>` tokens mark supervision points and the corresponding answers serve as intermediate (ancillary) supervision signals.

We evaluate *IPS* on a public Hate Speech Detection (HSD) dataset from MM-Soc (Jin et al., 2024), as well as two proprietary tasks: Unoriginal Content Classification (UCC) and Adult Nudity and Sexual Activity (ANSA) detection. Experimental results demonstrate substantial improvements in F1 score and recall across multiple precision thresholds compared to baseline MLLMs trained with standard end-to-end SFT. Ablation studies further confirm that both the inclusion of ancillary labels and the structured process supervision design are critical to these gains.

Scaling process supervision with MLLM-generated annotations. While process supervision improves performance, it introduces additional an-

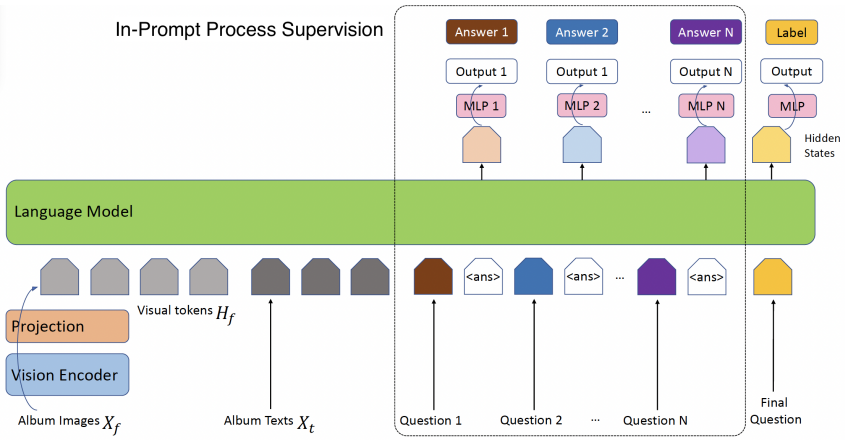

Figure 1: In-prompt Process Supervision (*IPS*) framework. The framework comprises a vision encoder, a modality-alignment projector, and a language model that jointly processes visual and textual tokens. During SFT, N questions are concatenated into a single prompt to improve both training and deployment efficiency. Each ancillary question terminates with an <ans> token; the hidden state corresponding to this token is passed through an MLP to predict the associated answer.

notation overhead beyond final-label supervision. Notably, ancillary labels serve as intermediate guidance rather than strict determinants of the final outcome, making them less sensitive to minor inaccuracies. Motivated by this observation, we explore automatically generating ancillary labels using general-purpose MLLMs.

Despite achieving only 75% agreement with human annotations, *IPS* remains robust to such noise. On the UCC task, replacing human-annotated process labels with MLLM-generated ones results in only marginal performance changes, demonstrating the scalability of our approach for large-scale industrial deployment.

2 Related Work

2.1 Multimodal Large Language Models (MLLMs)

Recent advancements in Multimodal Large Language Models (MLLMs), such as GPT-4V (Achiam et al., 2023), GPT-4o (Achiam et al., 2023), Gemini (Team et al., 2023), and Claude-3.5, demonstrate remarkable versatility across a wide range of vision tasks, including single-image, multi-image, and video analysis. Although many models are tailored for specific task types, the open-source model LLaVA-OneVision (Li et al., 2024) is designed for strong cross-task performance and exhibits robust feature transfer. Several other versatile models, such as Video-LLaMA (Zhang et al., 2023), and VILA (Lin et al., 2024), further highlight the growing potential of multi-scenario MLLMs. However, applying MLLMs in classification tasks remains challenging, as critical information for classification is not efficiently extracted by LLMs (Zhang et al., 2024; Li et al., 2025b; Sun et al., 2025; Li et al., 2025a; Meng et al., 2026; Chen et al., 2026).

2.2 Chain of Thought

Chain-of-Thought (CoT) prompting, introduced by Wei et al. (2022), has emerged as a powerful method to enable process supervision in LLMs. Leveraging structured prompting, CoT decomposes complex tasks into intermediate reasoning steps, allowing models to solve problems systematically. Building upon this foundation, Yao et al. (2024) and Long (2023) introduce a Tree-of-Thought (ToT) framework, a novel approach to multi-round question answering (QA) using language models. Besta et al. (2024) propose Graph-of-Thought (GoT) to model the reasoning process as a graph. In addition, fine-tuning with CoT (Yuan et al., 2023; Xiang et al., 2025; Wang et al., 2025) demonstrates better performance compared with direct SFT in reasoning tasks. Collectively, these studies highlight that leveraging contextual influence during the reasoning process can substantially enhance the performance of LLMs.

2.3 Process Supervision

Uesato et al. (2022) introduces a comparison between process supervision and outcome supervision in reasoning tasks and Lightman et al. (2023); Li et al. (2025c); Zhou et al. (2026) follow the study and shows that process-supervised reward

models (PRM) lead to a significant improvement in mathematical reasoning. Ma et al. (2023), transforms content moderation into a reasoning task and provides weak supervision in fine-tuning LLMs. Wang et al. (2024a) applies the PRMs to verify the generation of clinical notes.

3 Methodology

3.1 Problem Formulation

Given a video X and a set of violation labels L defined by a predefined policy, the content moderation model aims to identify the label in L that best characterizes the content, along with a confidence score. This score is subsequently used by the recommendation system to determine whether the content should be recommended.

3.2 Dataset Structure

Each policy is associated with a dedicated dataset annotated by professionally trained human reviewers in accordance with the corresponding policy documentation. In industrial content-moderation pipelines, such policies are typically operationalized as structured decision trees so that annotators can reach consistent judgments on complex cases. For each content instance X, the annotator follows a predefined decision tree T, producing a final label L_f along with a set of ancillary labels $(L_1, ..., L_m)$ corresponding to the intermediate decision nodes.

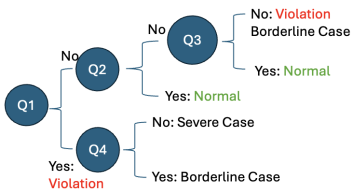

Figure 2: Illustration of the decision tree used by labeling agents. In practice, annotators may not strictly adhere to the tree in order to maintain flexibility, and only the final Normal/Guilty label undergoes formal quality assurance.

For each data entry (X, Y), the input is defined as $X = (X_f, X_t)$, where X_f denotes the visual frames and X_t represents the textual component of the content. The corresponding label is $Y = [L_f, [L_1, \ldots, L_m]]$. Here L_f is the final label to be predicted (typically binary), while $L_1, \ldots, L_m$ are ancillary labels derived from the intermediate nodes of the decision tree T (e.g., Figure 2). These ancillary labels either function as intermediate indicators that guide annotators toward the final decision or serve as fine-grained sub-labels (e.g., violation sub-types) that enrich the annotation. The number of ancillary labels m may vary even within the same policy, as certain branches of the decision tree can terminate early and some questions are optional.

Due to resource constraints, only the final label L_f undergoes formal quality assurance, while ancillary labels remain relatively noisy and are therefore often ignored in prior work. This underutilization represents a significant loss of valuable annotation signals. To address this issue, we propose *IPS*, a cost-efficient framework that effectively leverages these noisy ancillary labels to better align the model with policy requirements and improve final-label prediction performance.

3.3 *IPS*: In-Prompt Process Supervision

As illustrated in Figure 1, *IPS* incorporates multiple question prompts and their corresponding answers as process-level supervision signals during MLLM fine-tuning. Analogous to addressing simpler, related questions before tackling a complex one, this design simulates multi-step guidance within a single forward pass. It introduces negligible inference overhead, leverages previously unused ancillary labels, and improves the accuracy of the final prediction.

Model Architecture For each video frame or album image, the visual input X_f is first processed by a vision encoder $g(\cdot; \theta_g)$ to extract visual features:

$$Z_f = g(X_f; \theta_g). \tag{1}$$

The extracted features are then passed through a projector module $p(\cdot; \theta_p)$, which maps them into a shared multimodal representation space:

$$H_f = p(Z_f; \theta_p). \tag{2}$$

Here Z_f captures high-level visual representations, while H_f aligns them with the textual modality for joint modeling.

The language model $LM(\cdot; \theta_{\text{lm}})$ processes a multimodal token sequence of length L, consisting of visual tokens H_f, textual tokens X_t, and a policy-specific prompt P. The model typically adopts a decoder-only transformer architecture, which produces a sequence of hidden representations H corresponding to each input token:

$$H = LM([H_f, X_t, P]; \theta_{\text{lm}}). \tag{3}$$

***IPS* and Classification Layers** During SFT, N rounds of ancillary question-answer pairs are incorporated into the prompt P to provide process supervision:

$$P = \text{“}\{question1\}\texttt{<ans>}\{question2\}\texttt{<ans>} \ldots \{final_question\}\text{”} \tag{4}$$

Each ancillary question i is followed by a special <ans> token. Let d denote the hidden-state dimension of the language model. The hidden state corresponding to this token, $H^i_{<ans>} \in R^d$, is fed into an MLP classifier $f^i_{\text{cl}}(\cdot; \theta^i_{\text{cl}})$, which projects it into the answer space:

$$\hat{y}^i = f^i_{\text{cl}}(H^i_{<ans>}; \theta^i_{\text{cl}}), \tag{5}$$

where $\hat{y}^i \in R^{|\mathcal{Y}^i|}$ is the predicted distribution and $\mathcal{Y}^i$ denotes the set of possible classes for question i. The hidden representation of the final token in the sequence is passed through another MLP to produce the prediction for the final label.

During training, the model minimizes a weighted sum of the losses from all N ancillary questions and the final question:

$$\text{Training_Loss} = \sum_{i=1}^{N+1} w_i \cdot \mathcal{L}(\hat{y}^i, y^i). \tag{6}$$

The weights w_i are hyperparameters, with the default configuration:

$$w_i = \begin{cases} 0.1, & i = 1, \ldots, N \text{ (ancillary questions)} \\ 1, & i = N+1 \text{ (final question)} \end{cases} \tag{7}$$

Here, $\mathcal{L}(\cdot)$ denotes the loss function (e.g., cross-entropy). We experimented with ancillary weights in $\{0, 0.05, 0.1, 0.15, 0.2, 1\}$ and found that a stable default around 0.1 performs robustly across multiple tasks when $N < 5$. The fact that a single uniform weight works well across three heterogeneous tasks (UCC, ANSA, HSD) without per-task tuning indicates that the benefit of *IPS* comes from its structural design rather than careful hyperparameter calibration. In settings where certain ancillary criteria function as hard vetoes in the policy (e.g., strict keyword or entity gates), a hierarchy-aware or learnable weighting scheme could be a natural extension, which we leave to future work.

The core intuition behind *IPS* is twofold:

(1) *Additional Task-Specific Information*. Incorporating domain-tailored process labels provides richer supervision signals during SFT, enabling the model to better capture policy-relevant features.

(2) *Chain-of-Thought Reasoning in Decoder-Only Architectures*. By placing ancillary questions before the final question, the model accumulates intermediate supervision signals prior to making the final prediction. This structure aligns naturally with the Chain-of-Thought paradigm. Supervising earlier tokens encourages hidden states to encode meaningful intermediate semantics, improving reasoning coherence and final prediction accuracy. The decoder-only architecture facilitates this process, as each token attends to all preceding tokens, allowing process supervision to directly influence the final decision.

3.4 Process Annotation through MLLMs

We leverage general-purpose MLLMs, such as GPT-4o and LLaVA, to perform zero-shot process labeling. Because each ancillary question is intentionally designed to be clear and narrowly scoped, the risk of hallucination is substantially lower than when directly predicting final labels. Moreover, we empirically demonstrate that *IPS* remains robust under noisy process supervision. A detailed analysis of this robustness is provided in Section 5.1. This strategy enables *IPS* to scale seamlessly within the supervised fine-tuning (SFT) pipeline of MLLMs for large-scale deployment.

4 Experiments and Results

We conducted our experiments on two proprietary datasets, UCC (Unoriginal Content Classification), ANSA (Adult Nudity and Sexual Activity), and one open-source dataset, MM-Soc HSD (Hate Speech Detection benchmark for Multimodal Large Language Models in social media platforms) (Jin et al., 2024). Across all evaluated datasets, models incorporating the proposed *IPS* mechanism consistently outperform both baseline models and current state-of-the-art approaches, achieving top rankings on the leaderboard.

4.1 Experiment Setup

4.1.1 Datasets

The UCC task classifies posts as OC (Original Content) or UC (Unoriginal Content). The dataset includes $150K$ training and $8K$ test samples, each containing images with accompanying text. Ev-

ery sample has a binary final label (OC/UC) and four binary sub-issue labels that serve as process supervision.

The ANSA dataset is a large in-house benchmark for pornographic content detection, comprising 4M training and 21K test samples. It was originally annotated with a single binary label indicating the presence of adult material, nudity, or sexual activity (explicit or implicit), without process annotations. Therefore, all process labels in the ANSA experiments are generated by MLLM, using a pre-trained LLaVA-7b model as the annotator.

Furthermore, we evaluate the *IPS* method on the open-source HSD dataset from MM-Soc (Jin et al., 2024), which focuses on hateful meme detection. The dataset contains 8.5K training samples and 500 test samples. The three process labels used in this open-source dataset are generated with GPT-4o.

To show that *IPS* can work together with traditional CoT-incorporated training (Ma et al., 2023) and keep up with the latest MLLM models, we conduct experiments with the Qwen2.5-VL-7B (Team, 2025) model on the ANSA-borderline dataset, a subset of ANSA that introduces an even harder task of recognizing borderline content. This dataset consists of 200K training samples and 20K test samples. The CoT reasoning process in this dataset is generated with GPT-4o and the process supervision labels are provided by human labelers.

Details on model training are provided in Appendix E, while the specific prompts used for annotation can be found in Appendix G & H.

4.1.2 Metrics

To assess model performance, we primarily evaluate the recall at various precision levels, alongside the F1 score on both public benchmark and industrial datasets.

In addition, we deploy the best ANSA model online to build a content-based recommendation strategy and evaluate *IPS* via an A/B test using Sexual Suggestive View Rate (SSVR) and Inappropriate Content View Rate (ICVR)—the percentage of traffic viewing sexual or inappropriate content.

4.2 Key Results and Analyses

Table 1 shows that introducing the *IPS* mechanism improves performance across all three datasets evaluated in our experiments consistently. This indicates that the *IPS* method can be widely adopted for MLLM-based visual-language data classification tasks.

Task	Models	F1	R@P60	R@P65
UCC	SigLIP (Zhai et al., 2023)	-	70.6	62.0
	LLaVA-OV-0.5B Vanilla	66.7	71.6	67.5
	LLaVA-OV-0.5B *IPS*	**68.9**	**76.4**	**73.2**
	LLaVA-OV-7B Vanilla	68.9	75.4	72.2
	LLaVA-OV-7B *IPS*	**69.4**	**76.6**	**72.6**
ANSA	X-VLM (Bao et al., 2022)	45.7	38.0	32.4
	LLaVA-OV-0.5B Vanilla	54.9	49.9	44.1
	LLaVA-OV-0.5B *IPS*	**56.4**	**52.8**	**46.2**
	LLaVA-OV-7B Vanilla	56.6	52.0	46.6
	LLaVA-OV-7B *IPS*	**57.7**	**53.1**	**48.0**
HSD	LLaVA-1.5-7B	49.0	-	-
	LLaVA-1.5-13B	57.8	-	-
	LLaVA-OV-0.5B Vanilla	68.6	58.0	56.6
	LLaVA-OV-0.5B *IPS*	**68.9**	**60.4**	**59.2**
	LLaVA-OV-7B Vanilla	70.9	64.3	63.8
	LLaVA-OV-7B *IPS*	**72.6**	**69.1**	**67.6**

Table 1: Overall Performance (in %) comparison of Vanilla and *IPS* on all experiment datasets.

In the UCC task, *IPS* significantly improves offline recall over both the vanilla setting and the internal baseline SigLIP (Zhai et al., 2023) across various precision thresholds. These results highlight the effectiveness of *IPS* in enhancing model reasoning by aligning it with human-defined logic, leading to stronger performance across different model scales.

The ANSA dataset is the largest and most challenging in-house dataset, featuring a continuously updated test set that adapts to emerging trends. Despite the challenges posed by ANSA, the *IPS* model demonstrates substantial improvements over the X-VLM-based online model, outperforming all other in-house models and achieving the highest ranking on the internal leaderboard.

On the HSD dataset from MM-Soc (Jin et al., 2024), *IPS* again outperforms vanilla models in both F1 and precision-recall. Case studies further illustrate the interpretability and robustness of its sub-task predictions. See Appendix D for details.

5 Consistency Across Different Dataset Sizes

To assess the stability of *IPS*, we trained models on progressively larger subsets (854K, 1.7M, 2.5M, and 4M samples) in the ANSA task. As shown in **Table 5**, *IPS* consistently outperforms the vanilla model across all scales, indicating stable benefits under varying data availability. With the full dataset, our model achieves 63 precision at 50 recall, establishing a new state-of-the-art on the internal leaderboard. Gains are most pronounced in low-data settings, a trend also observed in the UCC task (**Table 2**)

Models	Data Size	F1	R@P60	R@P65	R@P70	R@P75	R@P80
LLaVA-OV-0.5b Vanilla	12k	51.7	35.1	29.7	23.5	18.3	13.7
LLaVA-OV-0.5b *IPS* (w/ MLLM data)		**56.1**	**48.6**	**39.5**	**32.0**	**24.8**	**18.1**
LLaVA-OV-0.5b *IPS* (w/ Human data)		55.5	45.3	36.1	29.8	20.2	15.6
LLaVA-OV-0.5b Vanilla	24k	56.2	50.8	46.1	41.4	31.7	24.3
LLaVA-OV-0.5b *IPS* (w/ MLLM data)		59.5	58.4	52.4	46.1	38.7	31.2
LLaVA-OV-0.5b *IPS* (w/ Human data)		**60.6**	**60.7**	**55.7**	**49.0**	**43.2**	**37.5**
LLaVA-OV-0.5b Vanilla	54k	62.1	64.0	56.6	52.2	44.5	37.6
LLaVA-OV-0.5b *IPS* (w/ MLLM data)		**64.7**	**68.6**	**64.1**	55.4	48.1	43.6
LLaVA-OV-0.5b *IPS* (w/ Human data)		64.4	66.5	63.0	**59.1**	**53.2**	**48.0**
LLaVA-OV-0.5b Vanilla	90k	65.7	70.5	65.9	58.4	52.4	46.5
LLaVA-OV-0.5b *IPS* (w/ MLLM data)		**66.8**	**72.7**	67.1	**62.9**	57.7	49.6
LLaVA-OV-0.5b *IPS* (w/ Human data)		66.3	72.1	**67.6**	62.6	**57.9**	**51.2**

Table 2: Impact of using MLLM-generated process label to replace human annotated label on UCC dataset. Different sizes of training data are compared to demonstrate the consistency of this approach. (Performance displayed in %)

5.1 Feasibility of Replacing Human-Annotated Process Labels with LLM-Generated Labels

To examine the feasibility of substituting human-provided process labels with those generated by MLLMs, we employed a general-purpose MLLM to produce labels for the UCC task, which already includes human annotations. We then trained models on subsets of varying sizes and compared the performance of *IPS* using MLLM-generated labels against models trained with human-labeled data.

As shown in **Table 2**, *IPS* trained on MLLM-generated data performs comparably to its human-labeled counterpart and consistently outperforms vanilla SFT across all data scales. These results demonstrate the robustness of *IPS*, even with imperfect supervision, and confirm its ability to leverage noisy but structured intermediate signals.

Overall, this highlights the potential of MLLMs as a reliable and scalable process supervisors within the *IPS* framework, offering a cost-effective alternative to manual annotation while maintaining strong performance in real-world settings.

Analysis of MLLM-generated Process Labels Although *IPS* is robust to imperfect process labels, assessing the quality of MLLM-generated annotations remains important. We manually reviewed 1,000 UCC samples, examining both process and final-label predictions to evaluate zero-shot process supervision.

As shown in **Table 3**, zero-shot predictions for the final question achieved an overall accuracy of 57.6%, with a missing rate of 12.23% due to ambiguous questions. This lower performance is expected, as general-purpose MLLMs lack domain-specific knowledge. In contrast, process questions—being simpler—showed no missing values and much higher human–machine agreement. These results suggest that while general-purpose MLLMs cannot yet match human annotators for final-label generation, they are suitable for generating process supervision labels that require less stringent quality verification. Prompt examples are provided in Appendix G.

Question	Agreement Rate
Final Question (Is the content UCC)	57.60
Process Question 1 (Watermark)	79.10
Process Question 2 (UGC)	66.95
Process Question 3 (Text originality)	74.29
Process Question 4 (Make a good serie)	77.68

Table 3: Zero-Shot Consistency with human-provided label (in %) of MLLM annotator on UCC task. We report a 'human–machine agreement rate' rather than absolute accuracy, since human-provided process labels lack formal quality assurance; only the final question's consistency can be meaningfully evaluated.

6 Deployment Results

Deploy *IPS* online for ANSA governance We conducted A/B experiments over a 14-day period, involving 10% of total traffic, yielding highly significant results ($p < 0.001$). *ANSA-IPS* is compared against a legacy X-VLM(Bao et al., 2022) based model. We evaluated the online metrics of inappropriate content view rate (ICVR) and sexual suggestive view rate (SSVR). Following the deployment of the upgraded model, ICVR decreased by 0.22% (95% CI: 0.22% ± 0.05%) and SSVR decreased by 1.3% (95% CI: 1.3% ± 0.08%). These metrics are critical for online model deployment, indicating that the *ANSA-IPS* model has an enhanced ability to reduce undesirable user experiences.

7 Conclusion

In this work, we present *IPS*, a novel framework for tackling challenging multimodal content moderation tasks. By integrating ancillary labels through a sequentially structured process supervision mechanism, *IPS* better aligns model reasoning with human labelers' chain-of-thought reasoning (*CoT*), achieving notable performance improvements over vanilla end-to-end SFT approaches. Replacing human-annotated ancillary labels with MLLM-generated ones yields only minor performance drops, demonstrating scalability while reducing manual labeling effort for real-world applications.

Ethical Considerations

Content moderation systems inherently involve trade-offs between over-moderation, which unfairly restricts legitimate expression, and under-moderation, which exposes users to harmful content. *IPS* is designed to be mindful of both ends of this trade-off. First, the live A/B evaluation in Section 6 shows that deploying *IPS* reduces the Inappropriate Content View Rate (ICVR) by 0.22% and the Sexual Suggestive View Rate (SSVR) by 1.3% on real user traffic, indicating that the method reduces harmful exposure without indiscriminate over-flagging. Second, beyond aggregate metrics, the interpretable ancillary predictions produced by *IPS* serve as an operational safeguard: instead of treating the final label as a black box, human moderators can audit the model's intermediate judgments on borderline cases and override the final decision when the ancillary signals disagree with policy-specific nuance. Finally, all proprietary training and evaluation data used in this work come from human-reviewed content-moderation pipelines operating under platform policies; no additional user data was collected for this study. The public HSD benchmark we evaluate on is released by Jin et al. (2024) under its original license, and we use it strictly within the terms of that release.

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*.

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

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, and 1 others. 2026. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*.

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

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024. Mm-soc: Benchmarking multimodal large language models in social media platforms. *Preprint*, arXiv:2402.14154.

Guangchen Lan, Huseyin A Inan, Sahar Abdelnabi, Janardhan Kulkarni, Lukas Wutschitz, Reza Shokri, Christopher G Brinton, and Robert Sim. 2025. Contextual integrity in LLMs via reasoning and reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3240–3251.

Yuqi Li, Qingqing Long, Yihang Zhou, Ran Zhang, Zhiyuan Ning, Zhihong Zhu, Yuanchun Zhou, Xuezhi Wang, and Meng Xiao. 2025b. Comae: Comprehensive attribute exploration for zero-shot hashing. *ICMR*.

Yuqi Li, Zijie Zhou, Zhiyuan Peng, Junhao Dong, Haochen You, Renye Yan, Shiping Wen, Yingli Tian, and Tingwen Huang. 2025c. A preference-driven methodology for efficient code generation. *IEEE Transactions on Artificial Intelligence*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *arXiv preprint arXiv:2310.03400*.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. 2026. Tri-subspaces disentanglement for multimodal sentiment analysis. *CVPR 2026*.

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. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 4872–4881, New York, NY, USA. Association for Computing Machinery.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2025. Qwen2.5-vl.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Hanyin Wang, Qiping Xu, Bolun Liu, Guleid Hussein, Hariprasad Korsapati, Mohamad El Labban, Kingsley Iheasirim, Mohamed Hassan, Gokhan Anil, Brian Bartlett, and 1 others. 2024a. Process-supervised reward models for clinical note generation: A scalable approach guided by domain expertise. *arXiv preprint arXiv:2412.12583*.

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*.

Zixuan Wang, Yu Sun, Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Dong, Zhuolin Hao, Hongyu Xiong, and Yang Song. 2025. Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1104–1112, Suzhou (China). Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. 2025. Promptsculptor: Multi-agent based text-to-image prompt optimization. *Preprint*, arXiv:2509.12446.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are visually-grounded language models bad at image classification? *Conference on Neural Information Processing Systems (NeurIPS)*.

Heng Zhou, Jing Tang, Jusheng Zhang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, and Jiawei Yao. 2026. Comem: Compositional concept-graph memory for vision–language adaptation. In *The Fourteenth International Conference on Learning Representations*.

A Ablation on *CoT* reasoning structure

To further assess the impact of the *CoT* reasoning structure in *IPS*, we conducted an ablation study on the UCC task using a multitask variant. In this variant, a multi-head MLP is applied to the final token's hidden state, with each head predicting labels for the four ancillary questions and the final question (**Figure 3**). This contrasts with *IPS*, which uses

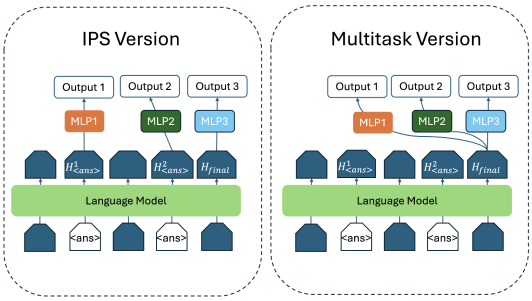

Figure 3: Difference between IPS and Multitask

MLPs on hidden states from distinct positions in the sequence. As shown in **Table 4**, while the multitask variant consistently outperforms the vanilla model, it does not match the performance of *IPS*.

The experiments suggest that the performance gains of *IPS* are attributable to two primary contributing factors:

- The incorporation of additional process supervision, quantified by the performance gap between the multitask and vanilla models (Multitask − Vanilla).
- The *CoT* reasoning structure. Process supervision on the carefully placed <ans> tokens effectively leverages this additional information, as evidenced by the gap between *IPS* and the multitask variant (IPS − Multitask).

Models	R@P60	R@P70	R@P80
LLaVA-OV-0.5B Vanilla	71.6	63.1	52.1
LLaVA-OV-0.5B Multitask	74.2	64.4	54.7
LLaVA-OV-0.5B *IPS*	**76.4**	**68.0**	**56.8**
LLaVA-OV-7B Vanilla	75.4	66.9	51.0
LLaVA-OV-7B Multitask	75.6	66.1	52.5
LLaVA-OV-7B *IPS*	**76.6**	**67.2**	**58.0**

Table 4: Performance (in %) Comparison on UCC task: MLLMs in Vanilla, Multitask, and *IPS* settings (LLaVA-OneVision 0.5B and 7B).

B *IPS*'s compatibility with autoregressive CoT training

Chain-of-thought (CoT)–instructed fine-tuning is a widely adopted technique in multimodal reasoning (Ma et al., 2023(Ma et al., 2023)). Compared with traditional CoT-centered fine-tuning, *IPS* focuses on a different aspect of improvement. While CoT-centered training emphasizes letting the model auto-regressively generate CoT sequences before producing an answer, *IPS* focuses on process-supervising a model within a fixed CoT-like input prompt.

In earlier experiments, we used randomly initialized, trainable MLP classifiers, which allowed us to compare the multi-task model with the *IPS* model. In this experiment, each MLP is instead initialized with the built-in lm_head from the original qwen2.5-vl checkpoint and kept frozen to ensure stability. The logits of selected tokens serve as the MLP outputs: for binary questions, the tokens are "No" and "Yes," and for multiple-choice questions, the tokens are "A," "B," "C," etc. This initialization preserves the model's text generation ability, thereby preserving full compatibility with CoT generation.

The ANSA-borderline dataset is a subset of the ANSA dataset. It contains ambiguous or borderline sexual content, providing a challenging testbed for reasoning-based classifiers. **Table 6** demonstrates that *IPS* can be applied in parallel with traditional CoT. Notably, combining Ma et al.'s CoT method with *IPS* achieves the latest state-of-the-art performance on the ANSA-borderline benchmark. This confirms that *IPS* enhances CoT-style training without interfering with the model's generative reasoning pipeline.

C Principles for Ancillary Question Generation

Although generating ancillary questions require domain knowledge and could be different for different issue, several empirical principles remained consistent throughout this project. In practice, the key consideration is **coverage**, defined as the proportion of samples answering "Yes" to a given ancillary question. Questions with moderate coverage (20-80% positive responses) provide informative supervision, whereas questions with coverage near 0% or 100% yield little to no informational gain. Furthermore, questions are preferred when their labels exhibit a strong correlation with the final target label.

During the development of the ANSA-IPS model, we designed 13 candidate ancillary questions and analyzed the distribution of "Yes" responses from the LLM annotator across negative (label 0) and positive (label 1) cases (**Figure 4**). Early questions, though almost exclusively "Yes" in positive cases, appeared in fewer than 5% of them, providing insufficient coverage. Later questions occurred in over 60% of positive cases despite "Yes" responses in about 30% of negative cases, offering a better balance between coverage and

Models	Data Size	F1	P@R50	R@P50	R@P55	R@P60	R@P65
LLaVA-OV-0.5b Vanilla	854k	48.6	46.5	46.7	41.8	37.1	29.6
LLaVA-OV-0.5b *IPS*		**53.0**	**55.2**	**54.9**	**50.2**	**45.2**	**40.3**
LLaVA-OV-0.5b Vanilla	1.7M	51.6	52.1	52.3	45.8	41.9	38.1
LLaVA-OV-0.5b *IPS*		**54.6**	**60.1**	**58.4**	**52.2**	**50.0**	**43.2**
LLaVA-OV-0.5b Vanilla	2.5M	53.2	55.4	54.8	51.0	46.5	41.0
LLaVA-OV-0.5b *IPS*		**55.6**	**60.4**	**60.0**	**55.4**	**51.0**	**43.0**
LLaVA-OV-0.5b Vanilla	4M	54.9	59.9	58.2	53.8	49.9	44.1
LLaVA-OV-0.5b *IPS*		**56.4**	**63.0**	**60.5**	**57.4**	**52.8**	**46.2**

Table 5: *IPS* with MLLM-generated process labels on ANSA dataset. Different sizes of training data are compared to demonstrate the consistency of this approach. (Performance displayed in %)

Task	Models	R@P50	F1
ANSA-borderline	Qwen2.5-vl-7B Vanilla	41.04	45.32
	Qwen2.5-vl-7B CoT	43.40	46.67
	Qwen2.5-vl-7B *IPS*	43.48	46.83
	Qwen2.5-vl-7B CoT + *IPS*	**45.38**	**47.66**

Table 6: ANSA-borderline classification Performance (in %) comparison of traditional CoT training, *IPS*, and combined.

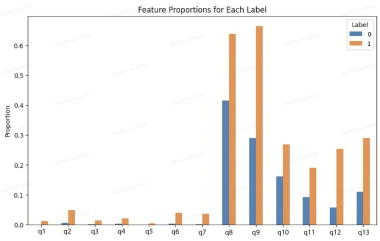

Figure 4: Positive Response Rate of Candidate Ancillary Questions. 13 candidate questions go through the first round of machine labeling in 1/5 of the dataset, evaluating it's coverage in non violation cases and violation cases respectively. Question example:"Is adult product mentioned in the content?". Details of the questions are omitted due to company privacy considerations.

discriminative power. Using sparse-response questions initially led to performance below the baseline, whereas replacing them with higher-coverage questions (8 and 9) substantially improved results (**Table 7**); these questions were ultimately used in **Table 5**.

For UCC and the ANSA-borderline dataset, we relied on human-annotated results for the initial

Task	Models	F1	R@P55
ANSA 1/5 training-set	Vanilla	48.6	41.8
	IPS with question 1-7	47.1	40.0
	IPS with question 1-2	47.5	40.9
	IPS with question 8-9	**53.0**	**50.2**

Table 7: Ablation (in %) on ancillary-question coverage on the ANSA 1/5 training-set. Low-coverage questions (1–7) can hurt performance, while high-coverage questions (8–9) substantially improve over the vanilla baseline.

attempt and did not create our own set of candidate questions. Instead, we aligned directly with the questions already present in the annotation template used by human agents. For the HSD task, our first attempt was simple and effective; the prompt details are provided in Appendix H.

D Case Study on HSD Dataset

From the case analysis in **Table 8**, *IPS* demonstrates more accurate judgment for complex cases, such as ID=1268 and 2783. This indicates that *IPS* facilitates the model's understanding of the policy by breaking down the problem.

Beyond aggregate accuracy, the per-dimension ancillary scores themselves serve as an interpretability trace that a free-text CoT output does not readily provide. In case 2783, for example, the vanilla model emits only a single low-confidence score (0.015) and offers no way to audit why the prediction was made, whereas *IPS* decomposes the decision into high ethnicity (0.980) and gender (0.926) signals with a near-zero religion signal (0.012), correctly surfacing the targeted group before producing the final label (0.991). In the clean case 1268, the low scores on all three ancillary dimensions give the moderator an immediate, structured justification for why the final prediction is safe. This structured decomposition also has a deployment-scale benefit: because every ancillary prediction is a discrete classifier output, policy teams can directly monitor what fraction of the video ecosystem triggers each individual ancillary criterion over time, enabling targeted policy refinement that free-text reasoning chains cannot easily support.

E Model Training Details

For MLLM, we use the LLaVA-OneVision model (Li et al., 2024) as the VL backbone, selecting its 0.5B and 7B versions for their balance be-

id	text	image	label	Vanilla score	Ethnicity score	Gender score	Religion score	*IPS*-final score
2783	a picture of my great great great grand grandfather's tractors they don't make them like they used to		1	0.015	0.980	0.926	0.012	0.991
1268	father and daughter bonding time is the best like and share if you agree		0	0.889	0.114	0.021	0.048	0.001
4538	bows here, bows there, bows everywhere		0	0.843	0.055	0.0168	0.006	0.009

Table 8: Comparison of Vanilla and *IPS* Scores on HSD Cases in MM-Soc Validation Set. *IPS* achieves more accurate predictions for complex problems by decomposing the issue.

tween computational efficiency and strong multimodal reasoning capabilities. For SFT, we adopt LoRA (Hu et al., 2021) with default parameters ($rank = 128, alpha = 256$) to optimize training efficiency. We trained the model with 16 A100-SXM-80GB GPU. The typical 5-epoch fine-tuning and evaluation process on UCC dataset takes 18 hours for 7b version and 10 hours for the 0.5b version.

To process video input, same number of frames are sampled from each video. All frames representations are converted to the 3d-Array format and uniformly resized to the same size.

After the fixed sized visual tokens, we designed a fixed m-token prompt for *IPS*, which is appended to the end of language input, facilitating structured process supervision. The text token from data sample is clipped to (256-m) token in order to avoid token overflow.

F Limitations

Our current evaluation focuses on domain-specific classification tasks, and we do not assess the text generation capabilities of the fine-tuned model. Exploring the generative potential of *IPS*-based models remains an avenue for future work.

While ancillary questions enhance the model's reasoning toward the final decision, designing effective ones requires domain expertise, as they must align closely with the underlying task. Poorly formulated or irrelevant questions may fail to provide useful guidance and could even impair performance. In addition, scaling this method to datasets without human-annotated process labels requires ancillary questions to be interpretable for general-purpose MLLMs. Empirical guidelines for constructing high-quality ancillary questions are provided in Appendix C.

Although *IPS* is designed to be noise-aware and tolerant of inaccuracies in ancillary labels, accurate final labels are still essential to ensure stable training and reliable evaluation.

G Prompt for UCC dataset with MLLM Process Annotation

The prompt consists of two parts:

1. Image-based Prompt: A list of image collections.

2. Text-based Prompt: A sequence of text-based questions.

G.1 Image-based Prompt

The image collection list includes 1 to 16 images, which are encoded using b64encode and then concatenated after the questions.

G.2 Text-based Prompt

1. Watermark presence. "'Watermark' is like '@username' from social media, not simple timestamp. Each image is considered as one image. Count the number of images with watermarks in the album."

2. Whether it is UGC (User-Generated Content). "UGC (User Generated Content) is considered as content is generated by regular users, such as selfies, artistic creations, life recordings, or concatenated images from online sources combined with self-created content. The opposite of UGC is PGC (Professionally Generated Content). PGC refers to content such as pictures of celebrities in entertainment/sports/politics, screenshots, posters, or coverage from TV series, movies, documentaries, and other platforms. Each image is considered as one image. Count the number of UGC images in the album."

3. Whether the image and the text title are relevant. "Original text is defined as content with emotional words (e.g., 'good,' 'happy,' 'disgusting') or symbols, subjective comments (e.g., 'I think the Doors are the best rock band'), or narrative storytelling (e.g., 'This movie tells the story of...'). Simple expressions without detail, like song lyrics or standalone sentences, are considered non-original. Determine if the given text is original."

4. Whether the image and the overall theme of the image collection are relevant. "Each image is considered as one image. Count the number of images whose content is related to the overall theme of the album."

H Prompt for MM-Soc Hate-speech Detection with MLLM Process Annotation

The prompt consists of two parts:

1. Image-based Prompt: one meme image.

2. Text-based Prompt: a sequence of three text-based questions answered within a single session.

H.1 Image-based Prompt

The image is encoded using `b64encode` and concatenated after the questions.

H.2 Text-based Prompt

The three ancillary questions are presented together in one session:

1. Ethnicity or Country. "Does the image and the given text contain satirical, discriminatory, harmful, cursing, racial, or other hateful content toward a certain ethnicity or country?"

2. Gender or a Certain Group of People. "Does the image and the given text contain satirical, discriminatory, harmful, cursing, racial, or other hateful content toward a certain gender or a certain group of people?"

3. Religion. "Does the image and the given text contain satirical, discriminatory, harmful, cursing, racial, or other hateful content toward a certain religion?"

Each question is answered independently with a binary response in $\{0, 1\}$, where 0 indicates no hateful content and 1 indicates hateful content in the targeted dimension.