Tool-MCoT: Tool Augmented Multimodal Chain-of-Thought for Content Safety Moderation

SHUTONG ZHANG, Stanford University, USA
DYLAN ZHOU, YINXIAO LIU, YANG YANG, HUIWEN LUO, and WENFEI ZOU, Google LLC, USA

The growth of online platforms and user content requires strong content moderation systems that can handle complex inputs from various media types. While large language models (LLMs) are effective, their high computational cost and latency present significant challenges for scalable deployment. To address this, we introduce Tool-MCoT, a small language model (SLM) fine-tuned for content safety moderation leveraging external framework. By training our model on tool-augmented chain-of-thought data generated by LLM, we demonstrate that the SLM can learn to effectively utilize these tools to improve its reasoning and decision-making. Our experiments show that the fine-tuned SLM achieves significant performance gains. Furthermore, we show that the model can learn to use these tools selectively, achieving a balance between moderation accuracy and inference efficiency by calling tools only when necessary.

CCS Concepts: • **Computing methodologies** → **Natural language generation**.

Additional Key Words and Phrases: Large Language Models, Agentic Framework, Reasoning

ACM Reference Format:
Shutong Zhang, Dylan Zhou, Yinxiao Liu, Yang Yang, Huiwen Luo, and Wenfei Zou. 2018. Tool-MCoT: Tool Augmented Multimodal Chain-of-Thought for Content Safety Moderation. In *Proceedings of 6th Workshop on Integrity in Social Networks and Media (Integrity 2026)*. ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The increasing popularity of online platforms and the explosion of user-generated content has created an urgent need for effective content moderation. However, harmful material is often expressed subtly through text, images, or their combinations, making traditional text- or image-only moderation systems insufficient. While LLMs are equipped with vision-language understanding and external tool integration can perform nuanced moderation, they are computationally expensive and difficult to deploy at scale. In contrast, SLMs are usually faster in terms of generation and require fewer resources to deploy, but they usually have limited ability to understand the complex relationship between modalities. This motivates a more efficient alternative: a small vision-language model that can reason agentically, leverage external tools to better understand the given image-text pair, then follow a structured chain-of-thought to make moderation decisions. In this paper, we present Tool-MCoT, a SLM that leverage external tools to obtain more information then perform further reasoning. By training these models on synthetic reasoning traces generated by LLMs, we show that the SLM is able to benefit from the tool information and learn to make tool-calling decisions from simple heuristics for content safety moderation via reasoning. We summarize our contributions as follows:

Authors' Contact Information: Shutong Zhang, tonyzst@stanford.edu, Stanford University, Palo Alto, California, USA; Dylan Zhou; Yinxiao Liu; Yang Yang; Huiwen Luo; Wenfei Zou, Google LLC, Mountain View, California, USA.

Manuscript submitted to ACM

(1) We propose an agentic tool framework consisting of OCR, image captioning, and object detection tools that the model can call to gather more information from the input image.
(2) We fine-tune an SLM that reasons over tool information for content safety moderation. We show that the fine-tuned model achieves performance gains on three open-source datasets.
(3) We further show that the SLM can learn the ability to use tools selectively on harder samples with simple tool calling heuristics, achieving a balance between accuracy and efficiency.

2 Related Work

Multimodal Content Moderation. Existing research focusing on multimodal content moderation mainly focuses on classification tasks, aiming to categorize content based on either image or video only [7, 11]. While these methods show promising results in identifying problematic content, most of them operate by combining features from different modalities or employing fusion [1, 21], lacking the ability to reason over text-image interactions. Our work distinguishes from the previous works by developing methods that enable reasoning across text and image content, moving beyond just the classification to address the reason behind the decisions.

Vision Language Model. Vision-Language Models (VLMs) have demonstrated strong abilities in tasks that involve understanding mixed visual and textual inputs, such as image captioning and visual question answering [4, 8, 23]. However, directly applying those general purpose large models for content safety moderation tasks is challenging since they are typically not optimized for this task and specific policies. As a result, leveraging them for moderation often requires extensive additional prompting. This results in high computational costs due to the increased input length and complex processing, and large latency due to the size of the model, making real-time or moderation at scale difficult. In this work, we address these limitations by utilizing a small VLM specifically designed and optimized for the content violation task, which enables cheaper and faster content moderation.

Tool Augmented Reasoning. The integration of tools has significantly improved the reasoning abilities of LLMs, allowing them to access more information and perform complex computations [6, 10, 16]. The incorporation of these external tools enables the models to go beyond their internal knowledge and address tasks that require additional information. However, the benefits of tool augmentation have mainly been observed with large models, as these larger models possess the capability to understand tool functionalities, formulate appropriate queries, and interpret tool outputs in the decision-making process [19]. In contrast, smaller models are typically weaker in reasoning and planning; they often struggle with tool selection, parameterization, and the effective incorporation of external information [3, 14, 20]. In this work, we will employ a small LLM that is specifically fine-tuned for tool use, thereby overcoming the drawbacks of existing small models.

3 Methods

Agentic Tool Framework. The Agentic Tool Framework is a collection of specialized tools designed to augment the capabilities of language models in performing content safety moderation tasks. Each tool addresses a specific challenge of multimodal content analysis. The framework contains the following three components:

(1) **Optical Character Recognizor (OCR):** This tool extracts text from images, enabling the LLM to process and reason about written content within visual media.
(2) **Image Captioner:** This tool generates a high-level textual summary of an image, providing the LLM with essential visual context for content moderation.

(3) **Object Detector:** This tool identifies and locates specific objects within an image, allowing the LLM to recognize potentially harmful objects.

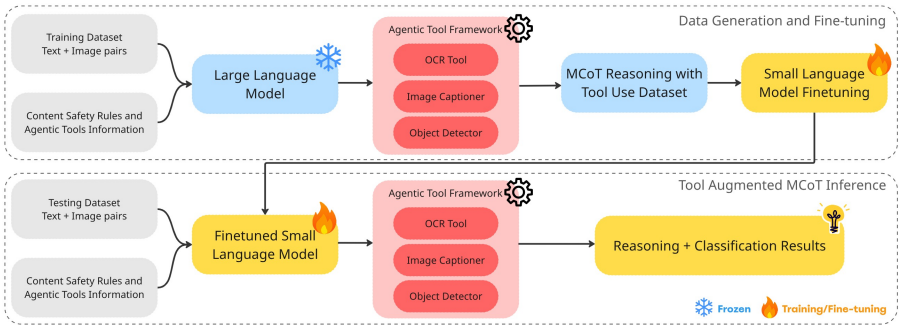

Fig. 1. **An overview of our two-stage pipeline.** We first generate tool reasoning data using a Gemini 2.0 flash as a teacher model, then use the generated reasoning data to fine-tune the SLM. During inference, the fine-tuned SLM utilizes the tool framework to conduct the content safety moderation task.

MCoT Data Generation. The first part of the pipeline involves using a LLM to generate Tool MCoT reasoning data to fine-tune the SLM. Specifically, we use the Gemini 2.0 Flash model [17] as the teacher model to generate the reasoning data. We prompt the Gemini model using the input image, text, tool information, and a format requirement to generate reasoning. Instead of following the STaR introduced in [22], we prompt the Gemini model to generate ten responses, setting the temperature and top_p to 1 to increase diversity. We then select samples that have at least one reasoning trace with a correct answer for our MCoT reasoning dataset. For samples with multiple correct reasoning traces, we randomly select one to use as training data for fine-tuning. For samples without any correct answer, we then use them as a part of the GRPO training dataset.

Small LM finetuning. The second stage of the pipeline involves using the previously generated MCoT Tool Use dataset to fine-tune a SLM. In this paper, we use the Gemma3-4B-IT model as the student model [18] We employ two fine-tuning approaches: one that enforces tool use for all samples, and another that trains the model to selectively use tools. In the enforced tool use setting, we force the model to use all tools by including all tool information within the prompt. The training process involves two stages: we first conduct LoRA tuning on the reasoning dataset generated by the teacher model [2], then further do GRPO using a combined format and answer reward [13]. In the selective tool use setting, the model learns to selectively call tools following a simple heuristics. It is trained to bypass tool use for simple samples that it can predict the correct result without any tool information. For more complex samples, the model is fine-tuned to call the OCR tool for images with overlaid text; call the object detection tool for images with complicated layouts (*i.e.*, has more than five detected objects); and always call image captioning tool. The SLM is then fine-tuned on a multi-turn dataset (as shown in Appendix C) that follows these rules.

Content Moderation. The final stage involves deploying the fine-tuned SLM for content moderation. During this phase, the SLM interacts with the Agentic Tool Framework to gather insights from multimodal content.

4 Experiments

Datasets. We test our proposed method on three multi-modal hate speech datasets. The MMHS150K dataset contains image-text pairs from Twitter, each manually labeled into one of six classes by three annotators [1]. We filtered for

samples with a majority label agreement, resulting in 74,177 training and 3,781 testing samples. The Hateful Memes dataset is a binary classification dataset contains text overlaid meme, each is hateful or not hateful [21]. It contains 8,500 samples in the training set and 1,500 samples in the testing set. The UnsafeBench dataset is a content safety multi-class image classification dataset [11]. It contains real-world images from the LAION-5B dataset [12] and AI-generated images from the Lexica dataset [15]. The dataset consists of 8,109 training samples and 2,037 testing samples.

Fine-tune SLM with enforced tool use. We first finetuned the small language model using tool reasoning data with LoRA tuning, then further train the model using GRPO. Experiment result shows that the model's performance improves on all three datasets after both LoRA tuning and GRPO.

	HatefulMemes		MMHS150K		UnsafeBench	
Model	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
Gemma3-4B-it	0.646	0.613	0.395	0.458	0.486	0.523
w/ LoRA	0.738	0.734	0.626	0.625	0.709	0.741
w/ Reasoning LoRA	0.777	0.763	0.655	**0.677**	0.743	0.757
w/ GRPO	**0.808**	**0.808**	**0.662**	0.674	**0.767**	**0.773**

Table 1. **Fine-tune small language model using tool reasoning dataset.**

Fine-tune SLM to selectively use tool. We further investigate whether the small language model is able to learn when and what tools to use. We finetune the model using tool conversation data then compare it with the base model and the tool reasoning LoRA-tuned model (Force Tool) on the MMHS150K dataset. Experiment results show that after fine-tuning the model to actively select tools, the tool calling ratio and run time (tested on a single Nvidia A100 GPU) reduced significantly while the performance remain almost unchanged.

Model	**Accuracy**	**F1-Score**	**OCR Ratio**	**Captioner Ratio**	**Detector Ratio**	**Run Time (s)**
Gemma3-4B-it	0.395	0.458	-	-	-	0.13
w/ Force Tool	**0.655**	**0.677**	1	1	1	1.67
w/ Select Tool	0.634	0.651	**0.203**	**0.484**	**0.263**	0.86

Table 2. **Finetune small language model to actively call tools when necessary.**

5 Conclusion

In this paper, we presented Tool-MCoT, an efficient and effective method for multimodal content moderation by augmenting a small language model with external tools. We demonstrate that a fine-tuned SLM can learn to leverage tool information to improve its reasoning and classification performance. We also show that the SLM can even learn to selectively use tools from simple tool calling heuristics to achieve a balance between performance and efficiency. By employing a pipeline that uses a large language model as a teacher to generate synthetic, tool-augmented reasoning data, we successfully transferred advanced reasoning capabilities to a much smaller, more resource-efficient model. Future work could explore methods such as using prompt optimization [9] to generate better prompts for the model, jointly training the language model's reasoning head with a classification head to improve inference efficiency, or investigating more advanced methodologies to enhance the SLM's selective tool-use capabilities beyond simple heuristics.

References

[1] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring Hate Speech Detection in Multimodal Publications. arXiv:1910.03814 [cs.CV] https://arxiv.org/abs/1910.03814

[2] 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:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[3] Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small Models Struggle to Learn from Strong Reasoners. *arXiv preprint arXiv:2502.12143* (2025).

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

[5] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

[6] Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. 2025. OctoTools: An Agentic Framework with Extensible Tools for Complex Reasoning. *arXiv preprint arXiv:2502.11271* (2025).

[7] 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. arXiv:2504.14904 [cs.SI] https://arxiv.org/abs/2504.14904

[8] Mohammad Ghiasvand Mohammadkhani, Saeedeh Momtazi, and Hamid Beigy. 2025. A Survey on Bridging VLMs and Synthetic Data. *Authorea Preprints* (2025).

[9] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. arXiv:2305.03495 [cs.CL] https://arxiv.org/abs/2305.03495

[10] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv:2307.16789 [cs.AI] https://arxiv.org/abs/2307.16789

[11] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. UnsafeBench: Benchmarking Image Safety Classifiers on Real-World and AI-Generated Images. arXiv:2405.03486 [cs.CR]

[12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV] https://arxiv.org/abs/2210.08402

[13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

[14] Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024. Small LLMs Are Weak Tool Learners: A Multi-LLM Agent. arXiv:2401.07324 [cs.AI] https://arxiv.org/abs/2401.07324

[15] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. 2024. Prompt Stealing Attacks Against Text-to-Image Generation Models. arXiv:2302.09923 [cs.CR] https://arxiv.org/abs/2302.09923

[16] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. ViperGPT: Visual Inference via Python Execution for Reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)* (2023).

[17] Gemini Team. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[18] Gemma Team. 2025. Gemma 3 Technical Report. arXiv:2503.19786 [cs.CL] https://arxiv.org/abs/2503.19786

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[20] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[21] Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, and Mei Chen. 2024. Rethinking Multimodal Content Moderation from an Asymmetric Angle with Mixed-modality . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[22] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. arXiv:2203.14465 [cs.LG] https://arxiv.org/abs/2203.14465

[23] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-Language Models for Vision Tasks: A Survey. arXiv:2304.00685 [cs.CV] https://arxiv.org/abs/2304.00685

A Training Setup

A.1 Hardware Selection

All experiments were conducted on a cluster of eight NVIDIA H100 (80GB) GPUs. Specifically, LoRA fine-tuning was performed using Distributed Data Parallel (DDP), while the GRPO training utilized DeepSpeed ZeRO Stage 3 to manage the increased memory overhead of group-based reinforcement learning.

A.2 LoRA Tuning Setup

We use the following hyperparameters for the LoRA tuning:

- **Optimizer**: AdamW [5] with cosine scheduler with 0.1 weight decay and warmup ratio
- **Data Precision**: torch.bfloat16
- **Learning Rate**: 1e-5
- **Epochs**: 2 for the MMHS150K Dataset, 3 for the HarefulMemes and UnsafeBench Dataset
- **Per GPU Batch Size**: 1
- **Gradient Accumulation Step**: 16
- **LoRA Module**: q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, down_proj
- **LoRA Rank**: 16
- **LoRA Alpha**: 32
- **LoRA Dropout**: 0.1

A.3 GRPO Setup

We use the following hyperparameters for the GRPO tuning:

- **Optimizer**: AdamW [5] with cosine scheduler with 0.1 weight decay and warmup ratio
- **Data Precision**: torch.bfloat16
- **Learning Rate**: 1e-5
- **Epochs**: 1e-6
- **Per GPU Batch Size**: 1
- **Gradient Accumulation Step**: 8
- **Loss Type**: dr_grpo
- **Num Iterations**: 2
- **Num Generations**: 16
- **Reward Type**: [answer reward, format reward]
- **Reward Weight**: [4, 1]
- **LoRA Module**: q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, down_proj
- **LoRA Rank**: 16
- **LoRA Alpha**: 32
- **LoRA Dropout**: 0.1

A.4 Inference Setup

We do inference using the following hyperparameters:

- **do_sample**: True
- **temperature**: 0.1
- **top_p**: 0.3
- **new_max_token**: 512

B Qualitative result

The qualitative examples shown in the image below demonstrate how tool information and internal reasoning improves model accuracy on the Hateful Memes dataset by enabling complex contextual grounding. Without reasoning, the model often relies on surface-level heuristics, leading to incorrect classifications of reclaimed slurs in positive contexts or failing to identify hateful memes that require historical knowledge, such as recognizing a Nazi crematorium. By utilizing information provided by the internal tool and the **<think>** process, it identifies the athletic setting in the first image and the historical significance of the crematorium in the second, allowing it to catch the nuance that a standard model misses.

Fig. 2. **Qualitative comparison between model outputs with and without tool reasoning.** The tool reasoning model (bottom) correctly identifies the nuanced context in both cases, where the standard model fails.

C Tool Calling Conversation

Figure 3 illustrates a flowchart-style diagram of a multi-turn chat conversation between a system instruction, a user, an AI model, and an external tool. It models how an AI uses reasoning tags (*<think>*) and function execution tags (*<tool>*, *<answer>*) to solve a content moderation task.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

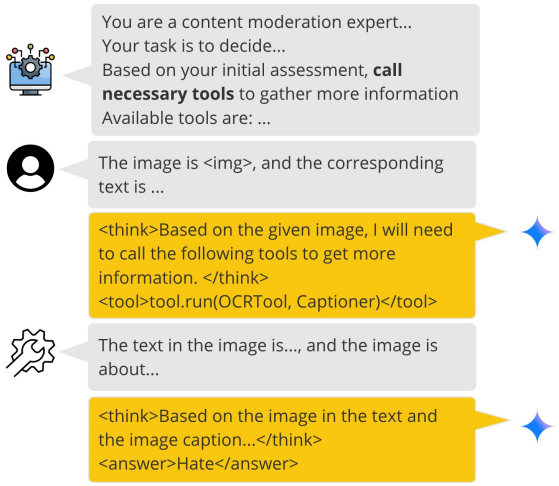

Fig. 3. **Multi-turn tool calling conversation.** Model select necessary tools for harder samples.