ICM-Assistant: Instruction-tuning Multimodal Large Language Models for Rule-based Explainable Image Content Moderation

Mengyang Wu[1,2,*†], **Yuzhi Zhao**[2,*‡], **Jialun Cao**[3], **Mingjie Xu**[4], **Zhongming Jiang**[4], **Xuehui Wang**[5], **Qinbin Li**[6], **Guangneng Hu**[7], **Shengchao Qin**[8,9], **Chi-Wing Fu**[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Huawei Hong Kong Research Center
[3]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
[4]Huawei Darwin Lab
[5]Artificial Intelligence Institute, Shanghai Jiao Tong University
[6]Department of Computer Science, National University of Singapore
[7]School of Computer Science and Technology, Xidian University
[8]Guangzhou Institute of Technology, Xidian University
[9]ICTT and ISN Laboratory, Xidian University
mywu@cse.cuhk.edu.hk; yzzhao2-c@my.cityu.edu.hk

Abstract

Controversial contents largely inundate the Internet, infringing various cultural norms and child protection standards. Traditional Image Content Moderation (ICM) models fall short in producing precise moderation decisions for diverse standards, while recent multimodal large language models (MLLMs), when adopted to general rule-based ICM, often produce classification and explanation results that are inconsistent with human moderators. Aiming at flexible, explainable, and accurate ICM, we design a novel rule-based dataset generation pipeline, decomposing concise human-defined rules and leveraging well-designed multi-stage prompts to enrich short explicit image annotations. Our ICM-Instruct dataset includes detailed moderation explanation and moderation Q-A pairs. Built upon it, we create our ICM-Assistant model in the framework of rule-based ICM, making it readily applicable in real practice. Our ICM-Assistant model demonstrates exceptional performance and flexibility. Specifically, it significantly outperforms existing approaches on various sources, improving both the moderation classification (36.8% on average) and moderation explanation quality (26.6% on average) consistently over existing MLLMs.

Caution: Content includes offensive language or images.

Code — https://github.com/zhaoyuzhi/ICM-Assistant

Introduction

In recent years, user-generated and AI-generated image contents (e.g., by DALL-E (Ramesh et al. 2021) or stable diffusion (Rombach et al. 2022)) keep multiplying on online platforms. Yet, some contents are against platform rules on child

---

*These authors contributed equally.

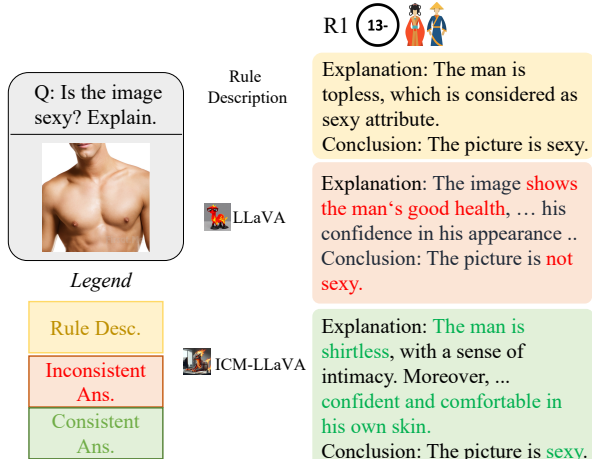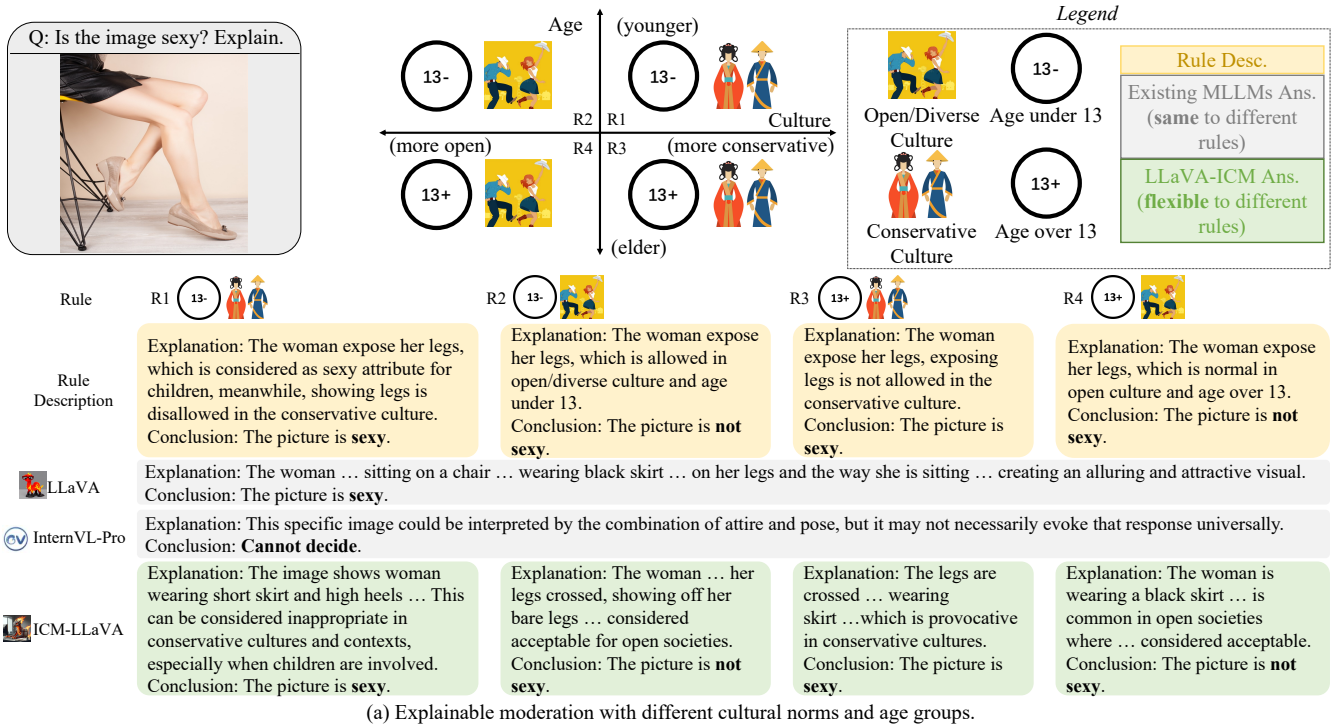†Internship with Huawei Hong Kong Research Center.

‡Corresponding Author and Project Lead.

protection (e.g., sexy, horrifying) (Facebook 2024; Search 2024; TikTok 2024) or on cultural regulations (De Gregorio and Stremlau 2023). Therefore, rule-based image content moderation has become increasingly significant.

Today, image content moderation faces three main challenges. First, moderation rules vary with cultures and ages. Different moderation strategies are needed for individual cultural regulations or child protection (see Fig. 1 (a), different combinations of culture and age have different moderation results on the same image). Second, transparency and openness receive increasing attention (Juneja, Rama Subramanian, and Mitra 2020). So, for the disclosure of review results, we need to provide reasons based on moderation rules, which increase the moderation complexity. Third, we need an in-depth understanding of the image contents to ensure accurate moderation results. Summarizing the challenges, there is a pressing need for *flexible*, *explainable* and *accurate* image content moderation (ICM) systems.

Traditional ICM systems (Son et al. 2023; Lees et al. 2022; Akyon and Temizel 2023; Momo et al. 2023) and datasets (Karavarsamis et al. 2013; Connie, Al-Shabi, and Goh 2017; NudeNet 2024; Phan et al. 2022) typically formulate ICM as a classification task. However, the image violation clues are often obscure and not identifiable solely by specific patterns or scenes. Therefore, traditional methods often obtain low classification accuracy. Moreover, their moderation processes are not explainable and the datasets are not flexible to different rules.

Recent multimodal large language models (MLLMs) (Liu et al. 2024; Bai et al. 2023; Team et al. 2023; Achiam et al. 2023; Markov et al. 2023) also perform ICM, by providing a general ability in describing and reasoning image contents. After practice, we find that existing MLLMs are not aligned with any specific moderation rules, resulting in two main inconsistencies with human moderators following various rules. (1) *Classification inconsistency*: the classification

Q: Is the image sexy? Explain.

Age (younger)

R2 | R1 — Culture
(more open) R4 | R3 (more conservative)
(elder)

13- / 13- / 13+ / 13+

Legend

Open/Diverse Culture — Age under 13
Conservative Culture — Age over 13

Rule Desc.
Existing MLLMs Ans. (**same** to different rules)
LLaVA-ICM Ans. (**flexible** to different rules)

Rule

R1 (13-)

Rule Description

R1: Explanation: The woman expose her legs, which is considered as sexy attribute for children, meanwhile, showing legs is disallowed in the conservative culture. Conclusion: The picture is **sexy**.

R2: Explanation: The woman expose her legs, which is allowed in open/diverse culture and age under 13. Conclusion: The picture is **not sexy**.

R3: Explanation: The woman expose her legs, exposing legs is not allowed in the conservative culture. Conclusion: The picture is **sexy**.

R4: Explanation: The woman expose her legs, which is normal in open culture and age over 13. Conclusion: The picture is **not sexy**.

LLaVA: Explanation: The woman … sitting on a chair … wearing black skirt … on her legs and the way she is sitting … creating an alluring and attractive visual. Conclusion: The picture is **sexy**.

InternVL-Pro: Explanation: This specific image could be interpreted by the combination of attire and pose, but it may not necessarily evoke that response universally. Conclusion: **Cannot decide**.

ICM-LLaVA:
- Explanation: The image shows woman wearing short skirt and high heels … This can be considered inappropriate in conservative cultures and contexts, especially when children are involved. Conclusion: The picture is **sexy**.
- Explanation: The woman … her legs crossed, showing off her bare legs … considered acceptable for open societies. Conclusion: The picture is **not sexy**.
- Explanation: The legs are crossed … wearing skirt …which is provocative in conservative cultures. Conclusion: The picture is **sexy**.
- Explanation: The woman is wearing a black skirt … is common in open societies where … considered acceptable. Conclusion: The picture is **not sexy**.

(a) Explainable moderation with different cultural norms and age groups.

Q: Is the image sexy? Explain.

R1 (13-)

Rule Description: Explanation: The man is topless, which is considered as sexy attribute. Conclusion: The picture is sexy.

LLaVA: Explanation: The image shows the man's good health, … his confidence in his appearance .. Conclusion: The picture is not sexy.

ICM-LLaVA: Explanation: The man is shirtless, with a sense of intimacy. Moreover, … confident and comfortable in his own skin. Conclusion: The picture is sexy.

Legend
Rule Desc.
Inconsistent Ans.
Consistent Ans.

(b) Classification and Explanation inconsistency.

ICM Accuracy (ICM acc.)

ICM acc.. +19.9%
ICM acc. +9.9%
ICM acc. +7.6%
ICM acc. +24.3%
ICM acc. +61.7%

MLLMs
LLaVA-v1.5 (7B)
ICM-LLaVA-v1.5 (7B)
LLaVA-v1.5 (13B)
ICM-LLaVA-v1.5 (13B)
mPLUG-Owl2 (7B)
ICM-mPLUG-Owl2 (7B)
Qwen-VL (7B)
ICM-Qwen-VL (7B)
CogVLM2 (19B)
ICM-CogVLM2 (19B)

Moderation Explanation Quality (MEQ)

(c) Results of moderation accuracy of diverse public MLLMs and our methods. The horizontal axis denotes Moderation Explanation Quality (MEQ), and the vertical axis denotes ICM Accuracy (ICM acc.). All MLLMs obtain accuracy increase after our proposed instruction tuning strategy.
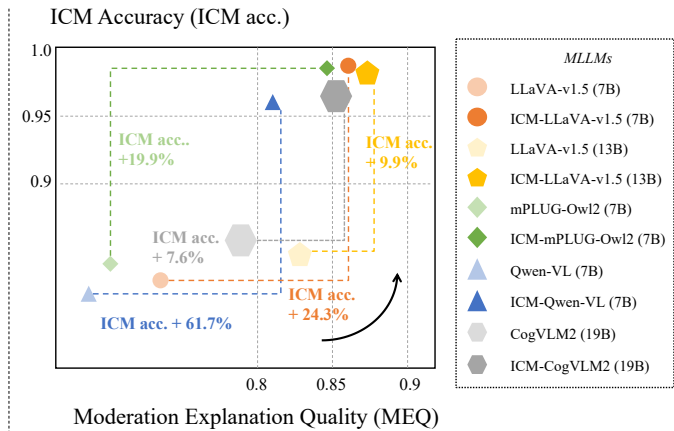
Figure 1: Overall framework: With input image and rules (on cultural norm and children protection), our method can (a) **flexibly** align with human moderators with four rules, and provide **explainable** results, overcoming the (b) classification and explanation inconsistency and achieving (c) more **accurate** moderation classification and explanations than the baseline MLLMs.

result/decision is inconsistent with the rules, and (2) *Explanation inconsistency*: the decision follows the rules but the explanation is wrong or incomplete, as shown in Fig. 1 (b). To address the issues, one solution is to instruction-tune MLLMs using domain-specific datasets. Yet, the lack of a multi-modal moderation dataset and the difficulty in customizing existing datasets into specific rules greatly limit the effectiveness and flexibility of the approach. To bridge

the gap, we need an effective ICM dataset and an efficient data generation pipeline, such that we can customize and enhance the moderation classification and explanation of MLLMs, and adopt it in the framework of rule-based ICM. We approach the goal by first designing a flexible rule-based ICM data generation pipeline. Given pre-defined moderation rules, we decompose it into several sub-categories, termed *attribute products*. Therefore, we can easily adapt to differ-
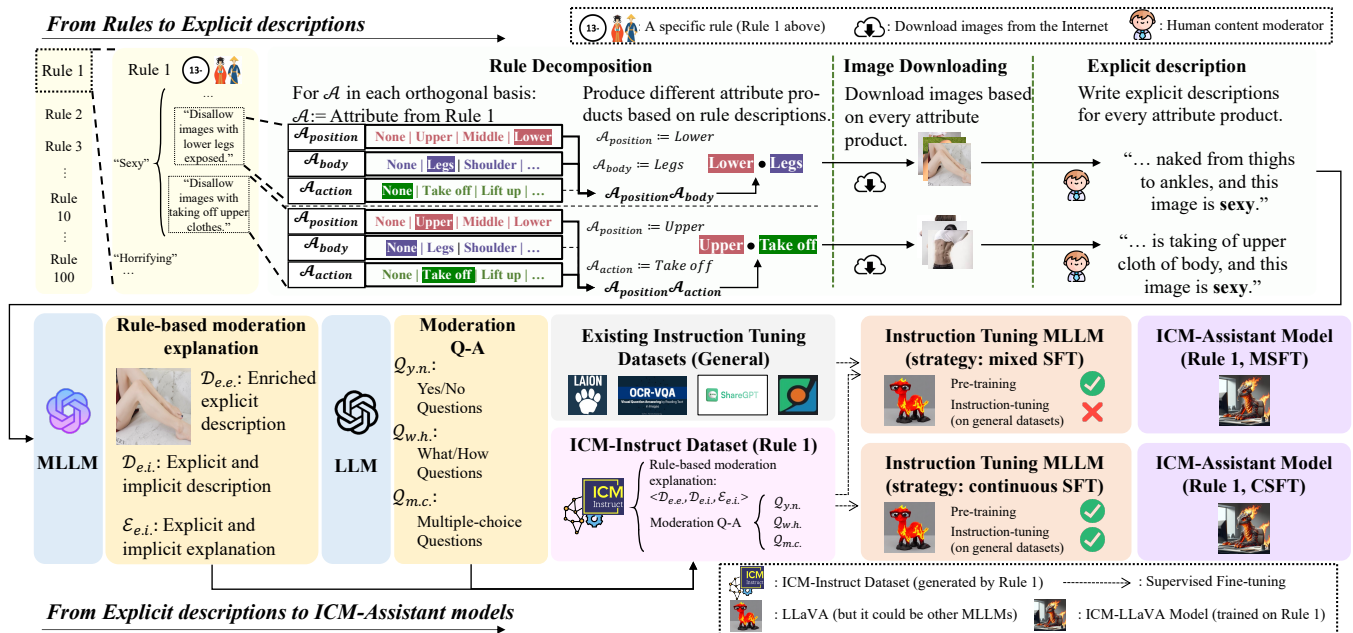
From Rules to Explicit descriptions

○13- 🧑: A specific rule (Rule 1 above) ⬇: Download images from the Internet 🧑: Human content moderator

Rule 1
Rule 2
Rule 3
Rule 10
Rule 100

Rule 1 ○13- 🧑

"Sexy" — "Disallow images with lower legs exposed." / "Disallow images with taking off upper clothes."

"Horrifying"

Rule Decomposition

For $\mathcal{A}$ in each orthogonal basis:
$\mathcal{A}$:= Attribute from Rule 1

$\mathcal{A}_{position}$: None | Upper | Middle | Lower
$\mathcal{A}_{body}$: None | Legs | Shoulder | …
$\mathcal{A}_{action}$: None | Take off | Lift up | …

$\mathcal{A}_{position}$: None | Upper | Middle | Lower
$\mathcal{A}_{body}$: None | Legs | Shoulder | …
$\mathcal{A}_{action}$: None | Take off | Lift up | …

Produce different attribute products based on rule descriptions.

$\mathcal{A}_{position} := Lower$
$\mathcal{A}_{body} := Legs$ — Lower • Legs
$\mathcal{A}_{position}\mathcal{A}_{body}$

$\mathcal{A}_{position} := Upper$
$\mathcal{A}_{action} := Take\ off$ — Upper • Take off
$\mathcal{A}_{position}\mathcal{A}_{action}$

Image Downloading

Download images based on every attribute product.

⬇ ⬇

Explicit description

Write explicit descriptions for every attribute product.

🧑 "... naked from thighs to ankles, and this image is **sexy**."

🧑 "... is taking of upper cloth of body, and this image is **sexy**."

MLLM

Rule-based moderation explanation

$\mathcal{D}_{e.e.}$: Enriched explicit description

$\mathcal{D}_{e.i.}$: Explicit and implicit description

$\mathcal{E}_{e.i.}$: Explicit and implicit explanation

LLM

Moderation Q-A

$\mathcal{Q}_{y.n.}$: Yes/No Questions

$\mathcal{Q}_{w.h.}$: What/How Questions

$\mathcal{Q}_{m.c.}$: Multiple-choice Questions

Existing Instruction Tuning Datasets (General)

LAION OCR-VQA ShareGPT

ICM-Instruct Dataset (Rule 1)

ICM Instruct

Rule-based moderation explanation: $<\mathcal{D}_{e.e.}, \mathcal{D}_{e.i.}, \mathcal{E}_{e.i.}>$
Moderation Q-A $\mathcal{Q}_{y.n.}$ $\mathcal{Q}_{w.h.}$ $\mathcal{Q}_{m.c.}$

Instruction Tuning MLLM (strategy: mixed SFT)
Pre-training ✓
Instruction-tuning (on general datasets) ✗

ICM-Assistant Model (Rule 1, MSFT)

Instruction Tuning MLLM (strategy: continuous SFT)
Pre-training ✓
Instruction-tuning (on general datasets) ✓

ICM-Assistant Model (Rule 1, CSFT)

ICM Instruct : ICM-Instruct Dataset (generated by Rule 1) --------> : Supervised Fine-tuning
🦙 : LLaVA (but it could be other MLLMs) : ICM-LLaVA Model (trained on Rule 1)

From Explicit descriptions to ICM-Assistant models

Figure 2: The overall pipeline from a specific set of moderation rules to a rule-based ICM-Assistant model. The first row illustrates rule decomposition, image downloading, and initialization of explicit descriptions. The second row presents the fully automatic data augmentation pipeline for ICM-Instruct dataset and the instruction-tuning process for ICM-Assistant models.

---

ent moderation rules by changing the classification labels of attribute products. Then, we adopt *multi-stage prompts* to progressively generate and enrich moderation explanations and Q-A pairs from the short annotations based on the "Chain-of-Thought" (CoT) (Wei et al. 2022), which decomposes ICM to several steps and makes ICM more accurately and interpretably. Doing so aligns with the ICM task, which requires the production of a moderation explanation as part of the moderation result. Accordingly, we construct a large-scale instruction-tuning **ICM-Instruct dataset**. With our ICM-Instruct dataset, we thereby perform instruction-tuning on various MLLMs by injecting content moderation reasoning knowledge. Results show the strong capabilities of our model and dataset in achieving flexible, explainable, and accurate moderation, as shown in Fig. 1 (c).

There are three technical contributions: (1) The flexible dataset generation pipeline based on concise moderation rules, short human annotations, and well-designed multi-stage prompts for data augmentation, largely surpassing existing moderation datasets; (2) The large-scale instruction-tuning ICM-Instruct dataset and an efficient evaluation benchmark for multiple ICM tasks; (3) The first method to accomplish diverse content moderation tasks, including moderation classification, explanation, and question-answering by adopting several instruction-tuning strategies to upgrade several MLLMs with ICM-Assistant. Results demonstrate that our method well moderates a wide range of images, outperforming existing MLLMs by an average of 33.7% in moderation classification and 26.6% in moderation explanation on various moderation tasks, rules, and data sources (e.g., user-generated, AI-generated, etc.).

Related Works

ICM datasets and methods. Image content moderation is an important topic of great practical need, given the proliferation of online platforms and social media. Common datasets include nudity/sexual content (Karavarsamis et al. 2013; Connie, Al-Shabi, and Goh 2017; NudeNet 2024; Phan et al. 2022), terrorism/violence (Schedi et al. 2015; Roy et al. 2017; Bianculli et al. 2020), etc. Method-wise, ICM can be single-modal (Son et al. 2023; Lees et al. 2022; Akyon and Temizel 2023; Momo et al. 2023; Zeng et al. 2024) or multimodal (Gupta et al. 2018; Tang et al. 2021; Yuan et al. 2024). While Large language models (LLMs) offer great help in data generation (Markov et al. 2023; Jin, Wanner, and Shvets 2024) and data augmentation (Ma et al. 2023), the resulting datasets and methods help mainly on classification, so they cannot meet the diverse content moderation requirements. Also, without accounting for the moderation process, incorrect moderation explanations are likely produced.

Multimodal Large Language Models. LLMs such as Llama 2 (Touvron et al. 2023), Mixtral (Jiang et al. 2024), ChatGPT (OpenAI 2024), and GPT4 (Achiam et al. 2023) demonstrate great achievements in recent years, by scaling (e.g., model/data size) a pre-trained language model to improve its capacity on downstream tasks. To extend the model abilities to visual tasks, many works combine visual and linguistic models for cross-modal comprehension, including well-developed closed-source models such as Gemini (Team et al. 2023) and GPT-4V (Achiam et al. 2023) and open-source models such as (Liu et al. 2024; Bai et al. 2023; Ye et al. 2023; Dong et al. 2024; Wang et al. 2023). Re-

cently, (Jha et al. 2024) explored a meme (images with text caption that is typically humorous or sarcastic) intervention method with an MLLM-based pipeline for understanding and intervening meme. Yet, it relies on text recognition and cannot provide results on moderation classifications and explanations following the established rules. Considering the strong scalability of open-source models, we build ICM-Assistants on them, e.g., LLaVA (Liu et al. 2024), mPLUG-Owl2 (Ye et al. 2023), and Qwen-VL (Bai et al. 2023).

Methodology

Problem Statement

As an image moderator under certain moderation rules, given an input image, existing neural-network-based methods normally regard such an ICM as a simple classification task. It can be formulated as maximizing a posteriori of the moderation classification result conditioned on the input and network parameters ϕ_r that follow moderation rules r:

$$\phi^* = \arg\max_{\phi} p(\mathcal{C}|I, \phi_r), \tag{1}$$

where I is the input image and $\mathcal{C}$ indicates the classification result on a moderation term, following pre-defined rules r. However, for explainable moderation and incorporating the explanation process in the moderation classification results, our ICM-Assistant outputs moderation results additionally with a moderation explanation $\mathcal{E}$, by optimizing

$$\Theta^* = \arg\max_{\Theta} p(\mathcal{E}|I, \Theta_r), \tag{2}$$

where Θ_r is the parameters of the ICM-Assistant with rules r and $\mathcal{E}$ is the moderation explanation including $< \mathcal{D}_{e.e.}, \mathcal{D}_{e.i.}, \mathcal{E}_{e.i.} >$, in which $\mathcal{D}_{e.e.}$, $\mathcal{E}_{e.i.}$, and $\mathcal{D}_{e.i.}$ denote the enriched explicit description, explicit-implicit description, and explicit-implicit explanation, respectively.

Based on this proposed model, we here cast the rule-based moderation task as a general pipeline, from data generation to MLLM instruction-tuning. As illustrated in Fig. 2, given sets of rules on a moderation term (e.g., "sexy"), we first initialize the explicit descriptions based on the decomposed rules and attributes product. Then, with such descriptions and moderation labels based on the rules, we enrich the explicit descriptions with MLLMs as moderation explanations and extract moderation Q-A with LLMs for a corresponding ICM-Instruct dataset. Finally, we instruction-tune the generated datasets and optionally the existing datasets for the corresponding ICM-Assistant models.

Rule Decomposition and Explication Descriptions

Observing the human moderators, they achieve proper moderation usually by first describing the images and then locating regions/semantics that violate the moderation rules. To reduce the workload and achieve flexible moderation, we start with a rule decomposition pipeline to split the rule items into short annotations for collecting images and writing explicit descriptions (one sentence or phrase related to the moderation rules) by moderators. Based on our observations, the rules defined for a moderation **term** (e.g., "sexy") can be decomposed into multiple orthogonal basis vectors,

terms as **attributes**. The values assigned to each orthogonal attribute are deliberately chosen to be distinctive in semantics (e.g., lower, middle, and higher in position), such that all the rule items can be represented as products of such attributes are distinct, termed **attributes products**.

To better explain, we use "sexy" as an example moderation term (Search 2024; Facebook 2024) to illustrate (Fig. 2 (top)): (1) images with sexy attraction content (e.g., a person's actions and clothing) are considered violations. Our method thereby decomposes the moderation rule items for "sexy" on the aspects of body bareness, actions of people, and camera focus position on the image. We then retrieve attribute products to summarize possible violations in the rule items, e.g, "lower leg" and "upper take off." Based on the attribute products, a dataset with real images is downloaded from the Internet by searching them on multiple search engines, including Google, Baidu, and Bing. A set of explicit descriptions paired with moderation classification results for each attribute product is also prepared by human moderators as well (see Fig. 2 (top right); in our implementation, there are 35 short sentences in total for the moderation term "sexy," following an example rule). More details can be found in *Supp. A*.

Rule-based Moderation Explanation Generation

With initial explicit descriptions and the paired images set prepared, we generate the moderation explanation for each image step by step, by enriching the shared explicit description with the assistance of current pre-trained MLLMs (Chen et al. 2023; Wang et al. 2023), (see Fig. 3 (a)): (1) We create a set of description-enrich prompts to enrich and include implicit descriptions from given explicit descriptions: enrich-explicit prompt: describe the explicit content of the image in detail; explicit-implicit prompt: describe the implicit content (e.g., overall atmosphere) of the image; explanation prompt: explain whether the image is sexy or not, based on the explicit and implicit contents). (see the details of prompts in *Supp. B*) (2) We use the above prompts as the input of pre-trained MLLMs and a tuple $< \mathcal{D}_{e.e.}, \mathcal{D}_{e.i.}, \mathcal{E}_{e.i.} >$. (enriched-explicit description, explicit-implicit description, explicit-implicit explanation) as moderation explanation $\mathcal{E}$. (3) We progressively increase the temperature parameters of MLLMs and repeat steps (1) and (2) to produce five samples per image. Based on images and several shared common explicit descriptions, our method can efficiently extends the amount, length, and variety of the moderation explanation data.

Moderation Q-A Generation

Though the explanation data comes with rich implicit and explicit information inside images, irrelevant content may be introduced in the previous stage. To encourage the model to focus on content directly related to the moderation terms and align it to follow a variety of instructions (Wu et al. 2023; Liu et al. 2024), we include additional data types, including *Yes/No* questions for binary answers, *What/How* questions for short-sentence answer, and *Multiple-Choice* questions for selected answers. With pre-trained LLMs (Touvron
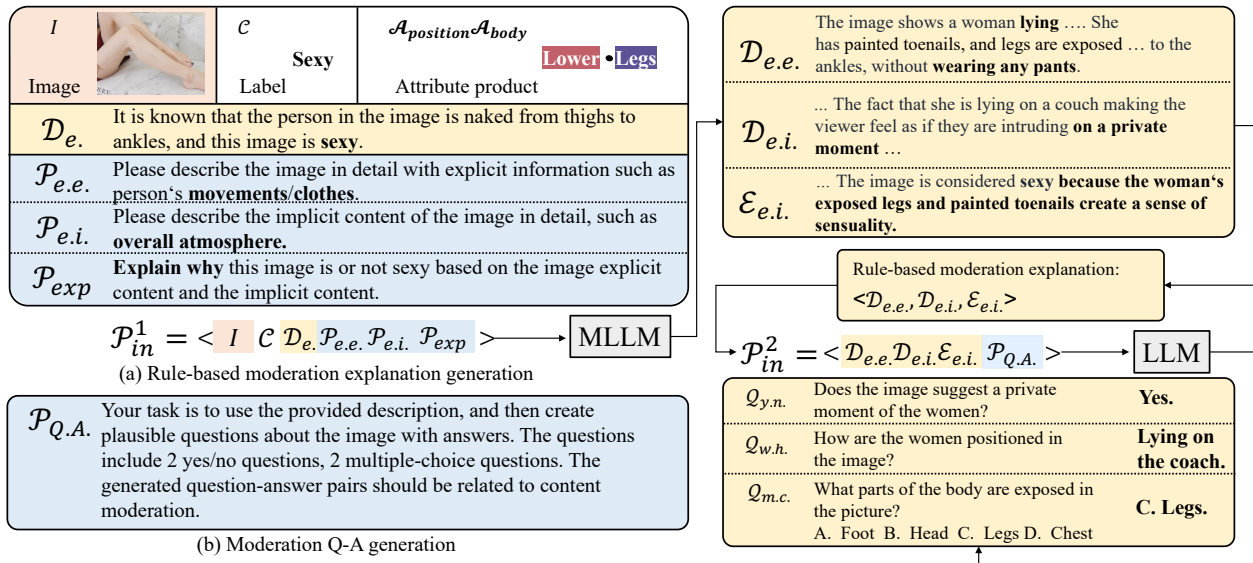
Figure 3: Pipeline for generating moderation explanations and Q-A pairs, from one attribute products "lower-leg".

et al. 2023; Bai et al. 2023), we thereby generate moderation Q-A from moderation explanation data in the following steps. (see Fig. 3 (b)): (1) For each image, we adopt the tuple $< \mathcal{D}_{e.e.}, \mathcal{D}_{e.i.}, \mathcal{E}_{e.i.} >$ (enriched-explicit description, explicit-implicit description, explicit-implicit explanation) as part of the prompt; and (2) We create Q-A prompts with the tuple and append tasks to generate plausible questions about the image with answers. In the prompts, we specify the question types and amounts, then force the output in table layouts (see *Supp. B*).

ICM-Instruct dataset

Due to the huge amount of possible rules and various moderation terms, we use the most common moderation term "sexy" (**S-ICM**) with two different rules **R1/R2** (**R1**: a more conservative rule that limits clothing style and skin exposure (Color expert 2022) for children under 13. **R2**: more open rules without clothing limitations but disallowing body positions with sexual attractions (MPA 2024) for the same age group) as a demonstration of the ICM-Instruct dataset. In summary, we collect 16K sexy images for training and validation, generate 75K tuples as the moderation explanation data, and create 246K moderation Q-A for the moderation question answering, following the rules R1/R2, termed the *ICM-Instruct* dataset.

The validation set includes 1,623 images, 1,623 classification labels, 379 moderation explanations from human moderators (a random subset of 1,623 images), and 17,701 moderation Q-A, based on rules R1/R2, termed separately as **ICM-Val-R1** and **ICM-Val-R2**. In addition, there are two test sets adopted in our experiments for further evaluation of the moderation ability for S-ICM, termed **ICM-Test**, which includes user-generated images and AI-generated images as follows: (1) **UGC**: 10,740 user-uploaded advertisement images downloaded from diverse advertisement platforms, where there are 1,443 not-sexy images, labeled according to

S-ICM R1. (See Fig. 4 (a).) We keep the data ratio in our test set to simulate real-world scenarios; (2) **AIGC**: 13,140 images generated by SDXL-1.0 (Podell et al. 2023). There are 1,000 initial SDXL-styled prompts automatically generated according to S-ICM R1. We generate 1,140 not-sexy images and 12,000 sexy images. (see Fig. 4 (b)) Human moderators have carefully checked all the above data samples and labels, and there are no privacy issues, as all the exposed real faces are blurred. Besides the ICM-Instruct for S-ICM, we collect relatively small sets for moderation terms "horrifying" and "gambling" for experiments. We provided more details in *Supp. A*.

Compared with the existing datasets (AIIA-PID4 (Karavarsamis et al. 2013), Adult content (Connie, Al-Shabi, and Goh 2017), NudeNet (NudeNet 2024), LSPD (Phan et al. 2022), and MemeGuard (Jha et al. 2024)) on ICM tasks (Table 1), our proposed dataset provides a significantly wider range of sub-categories (based on the attribute products) for moderation, a richer data format with moderation explanation and moderation Q-A for instruction tuning, and modalities that include normal images and meme.

Instruction-tuning strategies

In general, the training of open-source MLLMs includes two stages, (1) aligning the representation space of the visual backbone and the LLM with million-scale web data. (2) visual instruction tuning with a combination of human-labeled datasets. Considering the scale of the combination of all the datasets, we adopt a general strategy to mix its instruction-tuning datasets with the existing high-level datasets in the second stage, termed **MSFT** (mixed SFT), and a faster and more convenient strategy to add a continuous stage only with the ICM-Instruct after the original high-level tuning of MLLM, termed **CSFT** (continuous SFT). In our implementation, we train and validate various open-source MLLMs (e.g., LLaVA (Li et al. 2024) v1.5 (7B) and v1.6 (7B), Qwen-
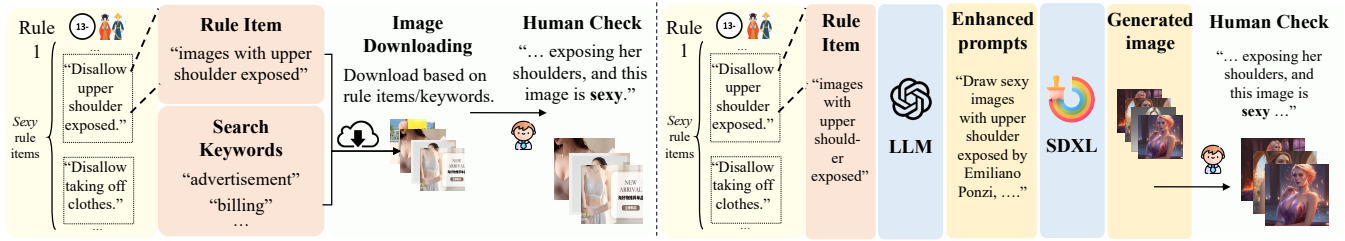
Figure 4: Illustration of the process for building ICM-Test dataset, including ICM-Test-UGC (left) and ICM-Test-AIGC (right).

Dataset	Categories	Modalities		Formats		
		I	M	C	E	Q
AIIA-PID4	4	✓		✓		
Adult content	2	✓		✓		
NudeNet	2	✓		✓		
LSPD	8	✓		✓		
MemeGuard	2		✓	✓	✓	
ICM-Instruct (ours)	45	✓	✓	✓	✓	✓

Table 1: Comparing our ICM-Instruct dataset with existing image (and video) content moderation datasets. In modalities, 'I' is for image, 'V' for video, and 'M' for meme (image with test caption covered). In formats, 'C' is for moderation classification, 'E' for moderation explanation, and 'Q' for moderation Q-A.

VL (Bai et al. 2023) (7B), mPLUG-Owl2 (Ye et al. 2023) (7B)) on the ICM-Instruct dataset.

Experiment

In this section, we present the experimental setting and implementation details and study the ICM abilities and ablations of our framework.

Test set. We employ both the ICM-Val and ICM-Test sets as the main test set. Additionally, to examine the generalizability of our model, we include the open-source NudeNet test set (NudeNet 2024) (we employ 4,658 neutral images and 13,858 porn images, which naturally follow our rule R1) along with our ICM-Test set for classification accuracy.

Metrics. We employ the following three metrics: (1) **ICM Accuracy (ICM acc.):** To better analyze the classification results, instead of directly adapting the outputs (i.e., yes or no) from MLLMs, we proposed an ICM-score to reflect the probability for MLLMs to output with an input prompt such as *"Is the image sexy?"*. By calculating the probability as the *softmax* of MLLMs output tokens (Wu et al. 2023). To calculate ICM-Accuary, we count the ICM-score larger than 0.5 as a positive reply and perform a binary comparison. (2) **Moderation explanation quality (MEQ)**: To measure the accuracy of the moderation explanation of MLLMs, we prompt MLLMs to answer the reasons following the classification result, e.g., *"Explain the reason."*. Inspired by (Fu et al. 2023), we applied LLMs (OpenAI 2024) to compare the reasons and moderation explanations labeled by human moderators according to precision, fluency, and comprehension (see our prompts in *Supp. C*). (3) **Moderation question answering accuracy (MQA acc.)**: As a pathway to moder-

Model	ICM-Val-R1	ICM-Val-R2
LLaVA-v1.5	0.794	0.680
ICM-LLaVA-v1.5-R1	**0.987**$_{+24.3\%}$	-
ICM-LLaVA-v1.5-R2	-	$0.971_{+42.8\%}$
LLaVA-v1.6	0.763	0.691
ICM-LLaVA-v1.6-R1	$0.986_{+29.2\%}$	-
ICM-LLaVA-v1.6-R2	-	**0.973**$_{+40.8\%}$
Qwen-VL	0.598	0.664
Qwen-VL-ICM-R1	$0.967_{+61.7\%}$	-
Qwen-VL-ICM-R2	-	$0.961_{+44.7\%}$
mPLUG-Owl2	0.821	0.735
mPLUG-Owl2-ICM-R1	$0.984_{+19.9\%}$	-
mPLUG-Owl2-ICM-R2	-	$0.962_{+30.9\%}$

Table 2: Results on ICM accuracy. Our ICM-assistant consistently brings signifcant improvement on four different MLLMs for both the ICM-Val sets on rules R1 and R2.

ation explanation, we evaluate the models' performance in answering moderation questions.

Implementation Details. All the experiments are conducted on 4x A100 80G GPUs with a default learning rate suggested by their official repositories. To save computing resources, we adopted CSFT in the main experiments.

Results on ICM Ability

Results on Flexibility on Rules. First, we present the results of the performance enhancement of our ICM-Assistant models trained with datasets constructed on different rules R1 and R2 on S-ICM. From Table. 2, we can see that with different rules, the accuracy on both ICM-Val-R1 and ICM-Val-R2 shows large improvements on all our models tuned with the corresponding rule-based dataset, with an improvement of 36.8% on average. From Fig. 5, we can see that with different rules on legs and belly, our models can conclude and explain consistently with the ground-truth rule descriptions provided by human moderators, indicating a large advantage of our models in flexibilities.

Moderation classification. We then experiment with the fundamental ICM classification tasks for the existing ICM methods and also the MLLMs on the Test set (ICM-Test and NudeNet (NudeNet 2024)). For quantitative comparison, Table 3 shows the comparison of ICM accuracy of existing open-source ICM methods, similar architectures trained on ICM-Instruct, pre-trained MLLMs, and ICM-Assistant
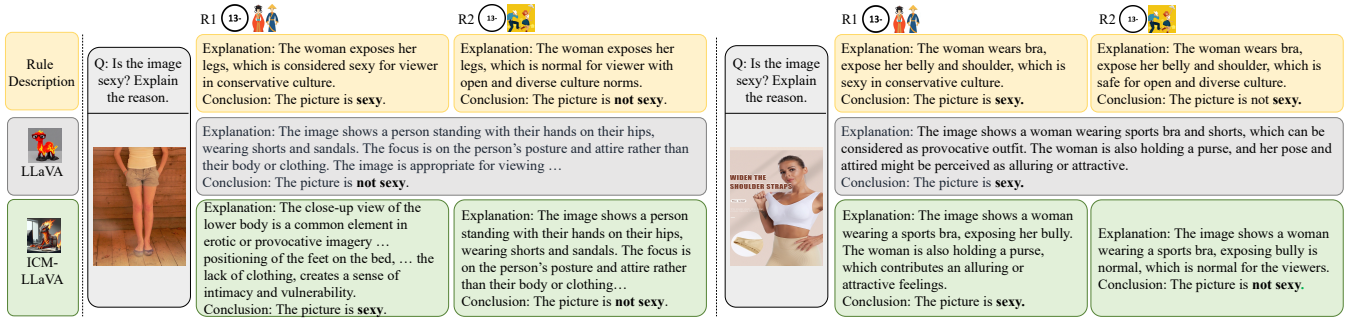
Figure 5: Comparison of ICM results with two different rules, R1 and R2. (See rule differences in Sec: ICM-Instruct dataset.)

Model	UGC	AIGC	Pornography
SD Safety Checker	0.040	0.307	0.589
NSFW model	0.089	0.679	0.637
NSFW detection	0.020	0.231	0.481
ICM-ResNet-152	0.566	0.902	0.764
ICM-DenseNet-201	0.586	0.889	0.767
ICM-CLIP-L/14-336	0.772	0.942	0.817
LLaVA-v1.5	0.896	0.940	0.923
ICM-LLaVA-v1.5	0.958	0.970	**0.983**
LLaVA-v1.6	0.929	0.957	0.923
ICM-LLaVA-v1.6	**0.972**	0.961	0.953
Qwen-VL	0.941	0.712	0.883
ICM-Qwen-VL	0.949	**0.974**	0.924
mPLUG-Owl2	0.828	0.953	0.916
ICM-mPLUG-Owl2	0.937	0.965	0.906

Table 3: Results on ICM accuracy by traditional methods and various MLLMs on the ICM-Test set.

Model	MEQ	MQA acc.
BLIP2	0.687	0.604
Kosmos-2	0.611	0.138
LLaVA-v1.5	0.696	0.801
ICM-LLaVA-v1.5	$0.844_{+21.26\%}$	$\mathbf{0.922}_{+15.11\%}$
LLaVA-v1.6	0.737	0.786
ICM-LLaVA-v1.6	$\mathbf{0.846}_{+14.79\%}$	$0.922_{+17.30\%}$
Qwen-VL	0.541	0.768
ICM-Qwen-VL	$0.789_{+45.84\%}$	$0.910_{+18.49\%}$
mPLUG-Owl2	0.666	0.767
ICM-mPLUG-Owl2	$0.831_{+24.77\%}$	$0.922_{+20.21\%}$

Table 4: Results on ICM explanation ability on the ICM-Val-R1 set.

Model	Horrifying	Gambling
LLaVA-v1.5	0.865	0.918
ICM-LLaVA-v1.5	$\mathbf{0.891}_{+3.01\%}$	$\mathbf{0.962}_{+4.79\%}$

Table 5: Accuracy of two zero-shot moderation terms.

MLLMs on the ICM-Test dataset. *First*, since open-source ICM methods (CompVis 2023; Laborde 2023; Falconsai 2024) are trained on datasets with limited moderation terms (e.g., pornography), it is hard for them to generalize to other related moderation terms (e.g., sexy). Therefore, their results on the UGC and AIGC test sets are unsatisfying. *Second*, trained with our ICM-Instruct dataset (as a classification task), several models with similar architectures (e.g., CLIP (Radford et al. 2021)) consistently achieve much better ICM accuracy on all test sets, indicating that the ICM-Instruct dataset helps the traditional models to improve generalization in real-world scenarios. *Third*, we also notice that open-source MLLMs (Li et al. 2024; Bai et al. 2023; Dong et al. 2024; Ye et al. 2023) have strong fundamental moderation ability on the three ICM-Test sets by outperforming most existing moderation methods. Trained with the full ICM-Instruct dataset, MLLMs consistently achieve much higher accuracy on most test sets.

Moderation explanation. Moderation explanation evaluation is based on MEQ and MQA acc. From Table 4, we notice the clear advantages of our ICM-Assitant models. Compared with the pre-trained MLLMs, our model achieves higher scores, with 26.6% higher in MEQ and 17.8% higher

in MQA acc. on average. Fig. 6 shows examples of moderation explanation of LLaVA-v1.5 and our ICM-LLaVA-v1.5. For LLaVA-v1.5, Fig. 6 (top) shows its inconsistency in both classification and moderation explanation results, by missing the possibility of unseen sexual acts. In Fig. 6 (bottom), though the image is considered to have a sexual act, the shape, appearance, and position of the orange and cherry are missed in deducting the reasons. Our method can provide consistent answers with those from human moderators and explain well. More moderation explanation and moderation Q-A results are presented in *Supp. C*.

Discussion on zero-shot ability. To study the generalizability of ICM-Assistants fine-tuned on S-ICM, we employ test sets for other moderation terms, "horrifying" and "gambling." Table 5 shows that models trained with one single term are also beneficial to other terms, revealing that MLLMs acquire the generalization ability of moderation with the assistance of our ICM-Instruct dataset. It shows the possibility that zero-shot abilities in moderation tasks could emerge with instruction tuning.
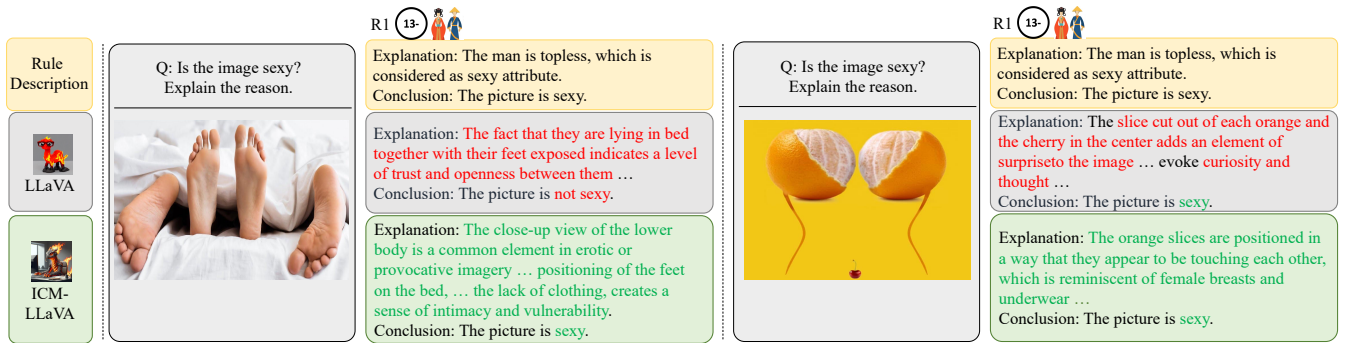
Figure 6: Comparing the moderation quality of ICM-LLaVA-v1.5 with its base model (marked in gray and yellow separately). Green texts indicate consistent replies, whereas red texts indicate inconsistent replies.

Strategies	ICM acc.	MEQ	L-B
LLaVA-v1.5	0.794	0.696	54.8
ICM-LLaVA-v1.5-MSFT	0.986	**0.856**	**56.0**
ICM-LLaVA-v1.5-CSFT	**0.987**	0.844	32.0

Table 6: Comparing the performance on different fine-tuning strategies. L-B denotes LLaVA-Bench.

Ablation Study

LLaVA-v1.5 (7B) is chosen as the baseline for ablations.

Instruction-tuning Strategies. Next, we study two SFT strategies on LLaVA-v1.5: CSFT and MSFT have similar performance in ICM acc. Though MSFT has the best explanation quality by mixing both data sources at the final training stage, CSFT consumes less time (6 hours vs 14 hours) with better ICM acc. results. To evaluate generalization ability, we include LLaVA-Bench (Li et al. 2024). Column 3 shows that MSFT models demonstrate better generalizability than the pre-trained LLaVA-v1.5, highlighting the potential of our ICM-Instruct dataset for tuning general tasks.

Components in our pipeline. Further, we provide experiments on the choice of MLLM and LLM in the data generation pipelines, as well as the prompts used in generating data with MLLMs. Compared with the baseline ICM-Instruct dataset generated with ShareGPT4V (7B) (Chen et al. 2023) and Llama 2 (70B) (Touvron et al. 2023), we choose CogVLM2 (19B) (Wang et al. 2023) and Qwen2 (72B) (Touvron et al. 2023) as alternatives. As shown in Table 7, our method achieves consistent performance on each combination, displaying only a small dependency on the choice of the component LLMs and MLLMs in the pipeline.

Data type of Moderation Q-A. Also, we analyze the contribution of data types of moderation Q-A in training the ICM-Assistants in Table 8, with the following findings: (1) All question types contribute to the ICM acc. (2) *M-C* questions contribute the most to the MQA acc. based on the results of experiments rows 1, 2, and 4. (3) The combination of *Y/N* and *W/H* questions result in the best MEQ, according to experiment row 4. *M-C* questions are not beneficial for generating moderation explanations with rich contents

MLLM	LLM	ICM acc.	MEQ	MQA acc.
C	Q	0.977	0.811	0.911
C	L	0.978	0.802	0.909
S	Q	0.984	0.838	0.915
S	L	**0.987**	**0.844**	**0.922**

Table 7: Performance of our ICM-Instruct with different choices of MLLMs and LLMs in generating moderation explanation and moderation Q-A. 'C' is for CogVLM2, 'S' for ShareGPT4V, 'Q' for Qwen2, and 'L' for Llama 2.

Y/N	W/H	M-C	ICM acc.	MEQ	MQA acc.
✓			0.973	0.844	0.835
	✓		0.978	0.841	0.810
		✓	0.895	0.831	0.894
✓	✓		0.984	**0.847**	0.830
✓		✓	0.986	0.841	0.920
	✓	✓	0.973	0.838	0.915
✓	✓	✓	**0.987**	0.844	**0.922**

Table 8: Performance of our ICM-Instruct with different types of moderation Q-A pairs (Y/N for Yes/No, W/H for What/How, and M-C for Multiple-Choice).

by restricting the answer as a single choice, yet resulting in higher ICM acc. and MQA acc.

Conclusion

In this work, we propose a flexible, explainable, and accurate ICM approach, by building a novel rule-based data generation pipeline, constructing the ICM-Instruct dataset, and instruct-tuning several MLLMs as ICM-Assistants for diverse ICM tasks. Experimental results show effective moderation of web images, user-generated ads, and AI-generated images under different rules. Notably, high accuracy in ICM tasks highlights our approach's potential in real-world applications, and zero-shot capabilities suggest strong generalizability in other moderation tasks.

Acknowledgments

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akyon, F. C.; and Temizel, A. 2023. State-of-the-Art in Nudity Classification: A Comparative Analysis. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bianculli, M.; Falcionelli, N.; Sernani, P.; Tomassini, S.; Contardo, P.; Lombardi, M.; and Dragoni, A. F. 2020. A dataset for automatic violence detection in videos. *Data in brief*, 33: 106587.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Color expert, P. b. C. E. 2022. Image censorship for the Middle East Market.

CompVis. 2023. Stable Diffusion Safety Checker. https://huggingface.co/CompVis/stable-diffusion-safety-checker.

Connie, T.; Al-Shabi, M.; and Goh, M. 2017. Smart content recognition from images using a mixture of convolutional neural networks. In *IT Convergence and Security 2017: Volume 1*, 11–18.

De Gregorio, G.; and Stremlau, N. 2023. Inequalities and content moderation. *Global Policy*, 14(5): 870–879.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; et al. 2024. InternLM-XComposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Facebook. 2024. Content policies for Facebook. https://transparency.fb.com/policies/community-standards.

Falconsai. 2024. NSFW Image Detection. https://huggingface.co/Falconsai/nsfw_image_detection.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. GPTScore: Evaluate as You Desire. *arXiv preprint arXiv:2302.04166*.

Gupta, D.; Sen, I.; Sachdeva, N.; Kumaraguru, P.; and Buduru, A. B. 2018. Empowering first responders through automated multimodal content moderation. In *2018 IEEE International Conference on Cognitive Computing (ICCC)*, 1–8.

Jha, P.; Jain, R.; Mandal, K.; Chadha, A.; Saha, S.; and Bhattacharyya, P. 2024. MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention. arXiv:2406.05344.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jin, Y.; Wanner, L.; and Shvets, A. 2024. GPT-HateCheck: Can LLMs Write Better Functional Tests for Hate Speech Detection? *arXiv preprint arXiv:2402.15238*.

Juneja, P.; Rama Subramanian, D.; and Mitra, T. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP): 1–35.

Karavarsamis, S.; Ntarmos, N.; Blekas, K.; and Pitas, I. 2013. Detecting pornographic images by localizing skin ROIs. *International Journal of Digital Crime and Forensics (IJDCF)*, 5(1): 39–53.

Laborde, G. 2023. Deep NN for NSFW Detection. https://github.com/GantMan/nsfw_model.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3197–3207.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ma, H.; Zhang, C.; Fu, H.; Zhao, P.; and Wu, B. 2023. Adapting Large Language Models for Content Moderation: Pitfalls in Data Engineering and Supervised Fine-tuning. *arXiv preprint arXiv:2310.03400*.

Markov, T.; Zhang, C.; Agarwal, S.; Nekoul, F. E.; Lee, T.; Adler, S.; Jiang, A.; and Weng, L. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15009–15018.

Momo, M. A.; Karim, H. A.; Sy, M. A.; Albunni, A.; Tan, M. J. T.; and AlDahoul, N. 2023. Evaluation of Convolution and Attention Networks for Nudity and Pornography Detection in Sketch Images. In *2023 IEEE Symposium on Computers & Informatics (ISCI)*, 7–12.

MPA. 2024.

NudeNet. 2024. NudeNet: lightweight Nudity detection. https://github.com/notAI-tech/NudeNet.

OpenAI. 2024. ChatGPT. https://openai.com/blog/chatgpt/.

Phan, D. D.; Nguyen, T. T.; Nguyen, Q. H.; Tran, H. L.; Nguyen, K. N. K.; and Vu, D. L. 2022. Lspd: A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15(1).

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Roy, A.; Paul, A.; Pirsiavash, H.; and Pan, S. 2017. Automated detection of substance use-related social media posts based on image and text analysis. In *2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI)*, 772–779.

Schedi, M.; Sjöberg, M.; Mironică, I.; Ionescu, B.; Quang, V. L.; Jiang, Y.-G.; and Demarty, C.-H. 2015. Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 1–6.

Search, G. 2024. Content policies for Google Search. https://support.google.com/websearch/answer/10622781?hl=en.

Son, D.; Lew, B.; Choi, K.; Baek, Y.; Choi, S.; Shin, B.; Ha, S.; and Chang, B. 2023. Reliable decision from multiple subtasks through threshold optimization: Content moderation in the wild. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 285–293.

Tang, T.; Wu, Y.; Wu, Y.; Yu, L.; and Li, Y. 2021. Video-moderator: A risk-aware framework for multimodal video moderation in e-commerce. *IEEE Transactions on Visualization and Computer Graphics*, 28(1): 846–856.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

TikTok. 2024. Content policies for TikTok. https://www.tiktok.com/legal/page/global/bc-policy/en.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2023. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. *arXiv preprint arXiv:2311.06783*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Yuan, J.; Yu, Y.; Mittal, G.; Hall, M.; Sajeev, S.; and Chen, M. 2024. Rethinking Multimodal Content Moderation from an Asymmetric Angle with Mixed-modality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8532–8542.

Zeng, W.; Liu, Y.; Mullins, R.; Peran, L.; Fernandez, J.; Harkous, H.; Narasimhan, K.; Proud, D.; Kumar, P.; Radharapu, B.; Sturman, O.; and Wahltinez, O. 2024. ShieldGemma: Generative AI Content Moderation Based on Gemma. arXiv:2407.21772.