

# Customize Multi-modal RAI Guardrails with Precedent-based predictions

Cheng-Fu Yang<sup>1†</sup>, Thanh Tran<sup>2</sup>, Christos Christodoulopoulos<sup>3†</sup>,  
Weitong Ruan<sup>2</sup>, Rahul Gupta<sup>2</sup>, Kai-Wei Chang<sup>1</sup>

<sup>1</sup>University of California, Los Angeles <sup>2</sup>Amazon AGI

<sup>3</sup>Information’s Commissioner’s Office  
cfyang@cs.ucla.edu

## Abstract

A multi-modal guardrail must effectively filter image content based on user-defined policies, identifying material that may be hateful, reinforce harmful stereotypes, contain explicit material, or spread misinformation. Deploying such guardrails in real-world applications, however, poses significant challenges. Users often require varied and highly customizable policies and typically cannot provide abundant examples for each custom policy. Consequently, an ideal guardrail should be scalable to the multiple policies and adaptable to evolving user standards with minimal retraining. Existing fine-tuning methods typically condition predictions on pre-defined policies, restricting their generalizability to new policies or necessitating extensive retraining to adapt. Conversely, training-free methods struggle with limited context lengths, making it difficult to incorporate all the policies comprehensively. To overcome these limitations, we propose to condition model’s judgment on “precedents”, which are the reasoning processes of prior data points similar to the given input. By leveraging precedents instead of fixed policies, our approach greatly enhances the flexibility and adaptability of the guardrail. In this paper, we introduce a critique-revise mechanism for collecting high-quality precedents and two strategies that utilize precedents for robust prediction. Experimental results demonstrate that our approach outperforms previous methods across both few-shot and full-dataset scenarios and exhibits superior generalization to novel policies.<sup>1</sup>

## 1 Introduction

Responsible AI (RAI) (Amazon, 2024) aims to develop AI systems that are safe, fair, helpful, and interpretable. Guardrails (Gehman et al., 2020; Welbl et al., 2021) are approaches designed to realize this goal, it can help ensure that AI behavior aligns with societal values, ethics, and the specific needs of diverse users. As AI systems increasingly interact with diverse types of media, the need for multi-modal guardrails becomes crucial. This system is required to filter out image contents that may promote hate and violence, reinforce harmful stereotype, contain sexual material, or propagate misleading or fraudulent information (Qu et al., 2024; Crone et al., 2018; Liu et al., 2023; Wu et al., 2020; Helff et al., 2024).

Different AI service/model providers often require different policies to regulate their AI systems (Dong et al., 2024b): one may prioritize filtering out erotic content, while another may be more concerned with violent material. Furthermore, even within the same category, definitions of what content should be blocked can vary significantly due to cultural differences—certain gestures or expressions might be considered offensive in one culture but neutral in another. An effective RAI guardrail system should, therefore, be both customizable and transparent. Not only should it accommodate user-defined policies, but it

---

<sup>†</sup>Work was done at Amazon.

<sup>1</sup>The code is available at:<https://github.com/joeyy5588/Customize-Guardrail>.

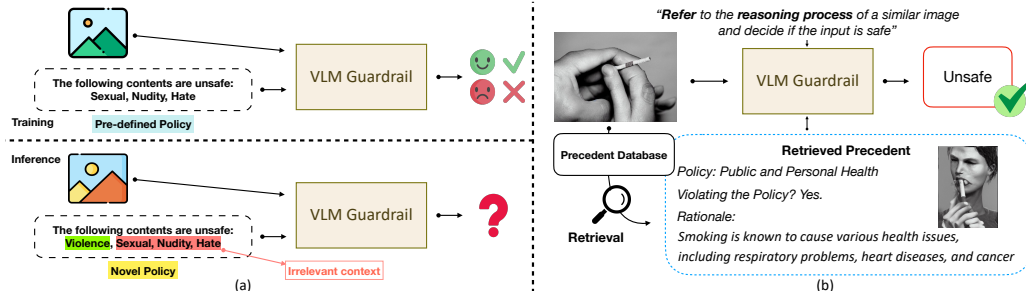


Figure 1: (a) Models fine-tuned on pre-defined policies struggle with novel policies and require extensive retraining to adapt. As the number of policy scales, during inference it becomes infeasible to include all the policies and may include irrelevant ones. (b) Our proposed method addresses these limitations by conditioning on structured *precedents*, which explicitly capture the model’s reasoning processes from past decisions. This enables the guardrail to flexibly adapt and accurately handle dynamically evolving and novel policies without extensive retraining.

should also provide clear explanations for each filtering decision, enhancing user trust and accountability.

However, a particular challenge arises from the need for flexibility in incorporating new rules, which often come in the form of few-shot learning tasks (Dong et al., 2024a). When users introduce new, specific filtering requirements, there is typically little data available to define these unique policies. As shown in Fig. 1 (a), prior fine-tuning methods (Schramowski et al., 2022; Qu et al., 2023; Helff et al., 2024; Schuhmann, 2023; Praneeth, 2019), become less feasible as they are pre-trained on pre-defined policies and involve substantial computational overhead for retraining with each new policy. While in-context learning methods (Dwivedi et al., 2023; Oba et al., 2023) offers an alternative, it becomes less viable as the number of policies grows. Limited context length constrains the ability to include all policies in the prompt, and including examples from irrelevant policies may degrade the guardrail’s effectiveness. Consequently, there is a need for methods that support scalable, flexible, and efficient customization of RAI guardrails without excessive retraining or labeling demands.

To address the above challenges, we propose to condition model’s prediction on a *precedent* database, as shown in Fig. 1 (b). A *precedent* is a structured example of model’s reasoning process, capturing how it previously predicts whether an image is policy-violating (PV) according to a specified policy and the rationale behind its decision. As model learns to reason with precedents, which capture contextually relevant nuances that static policy definitions may fail to encapsulate, it can naturally evolve and expand to accommodate new rules without extensive retraining.

In this paper, we first introduce a critique-revise mechanism to collect high-quality precedents, where the model reviews its own analyses, critiques incorrect predictions, and revises them accordingly. Importantly, this entire process operates without human in the loop. To utilize the precedent database effectively, we introduce two complementary methods that can be applied at training and inference time, respectively. First, for users with access to a local model like LLaVA (Liu et al., 2024), we demonstrate that fine-tuning with precedents collected via the critique-revise mechanism yields significant performance gains. Second, at inference time, the database can be leveraged through retrieval-augmented generation (RAG) to retrieve the most relevant precedent. This allows our method to scale efficiently even with a large number of diverse policies, and applicable for proprietary models like GPT-4V (Achiam et al., 2023) and Claude (Anthropic, 2023).

We evaluate our method on UnsafeBench (Qu et al., 2024), which encompasses 11 distinct RAI policies. Compared to prior baselines, our approach achieves substantial improvements of 6.6% and 6.8% in F1 scores across few-shot and full-dataset settings, respectively. Moreover, when we isolate each policy from the training set and treat it as a novel category—simulating the model’s generalization capability to unseen policies—our method outperforms existing approaches by 16.7% in terms of F1. These results underscore our

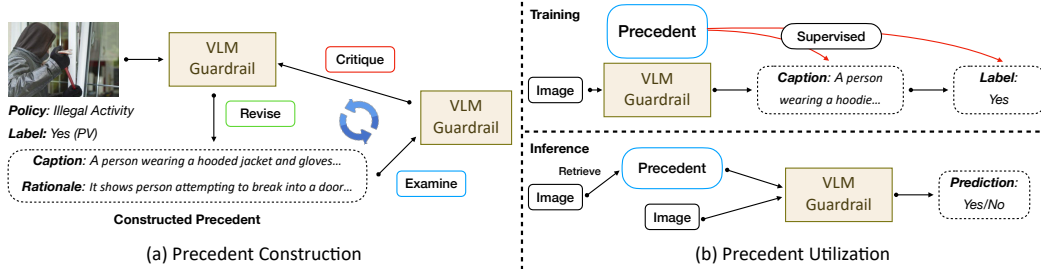


Figure 2: An overview of our framework leveraging a Visual-Language Model (VLM) with precedents. (a) During precedent construction, an instance is created through a critique-revise cycle where the VLM generates captions and rationales for images, receives critique, and revises its outputs. This enables our method to collect samples that were initially misclassified. (b) During training, the VLM is fine-tuned with the constructed precedents through supervised learning to enhance performance. At inference, our method retrieves the most relevant precedents for each input image, aiding in accurate policy-violation prediction with contextual support.

method’s effectiveness in efficiently scaling customizable RAI guardrails across diverse and dynamically evolving policy landscapes.

## 2 Related Work

**Datasets for Policy-Violating Visual Content.** Several efforts have been made to evaluate the effectiveness of current models in detecting policy-violating (PV) visual content. For instance, Wu et al. (2020) introduced a video dataset specifically for violence detection, while Lee et al. (2024) benchmarked text-to-image (T2I) models to assess issues related to gender, fairness, and toxicity. Jha et al. (2024) proposed methods to evaluate nationality-based stereotypes in T2I models, highlighting biases that may arise across cultural representations. In the realm of multi-modal large language models (MLLM), Liu et al. (2023) examined the resilience of these models against malicious attacks. Meanwhile, OpenAI (OpenAI, content no longer publicly available) established 11 categories for categorizing unsafe images, providing a foundational taxonomy for PV content detection. Building on this taxonomy, UnsafeBench (Qu et al., 2024) gathered an extensive dataset, which includes both real and synthetic images, thereby offering a comprehensive resource for evaluating model performance across diverse RAI policies.

**Current Multi-modal RAI Guardrails.** To filter out content for a given RAI policy, the most common approach is to fine-tune a classifier on the collected PV data. For example, NSFW Detector (Schuhmann, 2023) is trained on a subset of LAION-400M (Schuhmann et al., 2021) to detect explicit content, while NudeNet (Praneeth, 2019) leverages the YOLO backbone (Redmon, 2016) to build a nudity detection model. Q16 (Schramowski et al., 2022) is a binary CLIP (Radford et al., 2021) classifier that predicts the morality of the given image. Similarly, MultiHeaded (Qu et al., 2023) applies linear probing to fine-tune separate classification heads for five different RAI policies. While these methods demonstrate effectiveness in detecting specific types of PV content, they are typically limited to fixed policies and require extensive labeled data to build. While in-context learning has not been extensively explored for multi-modal guardrails, it has proven effective in reducing biases in language model outputs (Dwivedi et al., 2023; Oba et al., 2023). However, as the number of RAI policies increases, studies (Shi et al., 2023) have found that including irrelevant context in prompts can significantly degrade model performance.

## 3 Method

For the sake of clarity, we first define the task’s setting and notations. Specifically, the users can customize their RAI guardrail with  $P$  different policies, with each policy  $p$  contains  $N$ -shot data points. Each data point includes an image  $x$  and a binary label  $l$ , indicating whether the image is policy-violating (PV) or not. In Sec. 3.1, we first introduce the concept

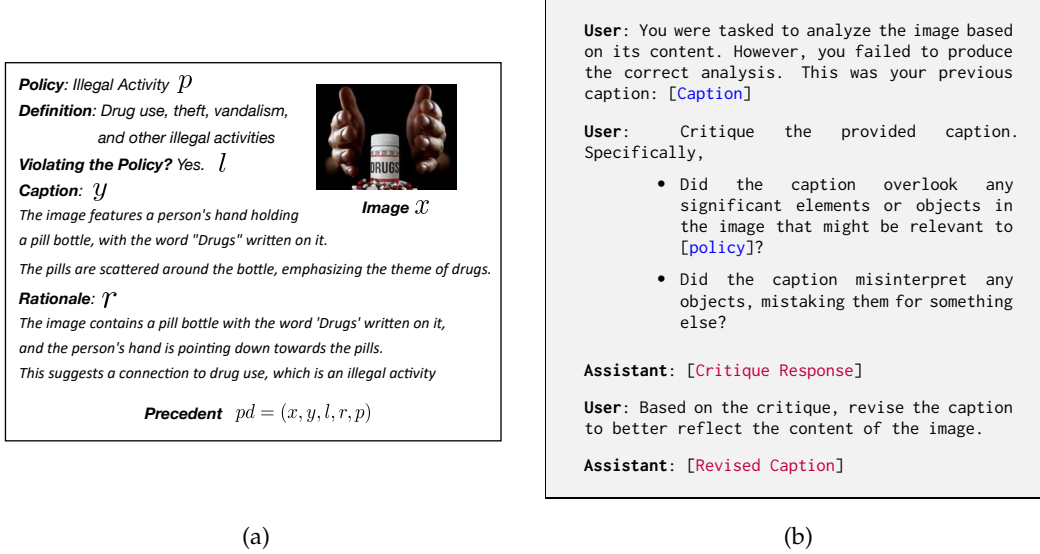


Figure 3: (a) An illustration of a precedent. A precedent is composed of a policy and the corresponding definition, an evaluation of whether the given image violates the policy, a caption of the image, and a rationale explaining the decision. (b) Prompt template used to guide the model in analyzing, critiquing, and revising image captions based on the given policies.

of *precedent* and how to utilize the limited user inputs to collect high-quality precedents. Then, in Sec. 3.2, we demonstrate how the constructed precedent database can enhance model’s performance. During training, fine-tuning with the collected precedents enhances the model’s robustness against novel instances, while at inference time, retrieval-augmented generation (RAG) allows the model to reference the most relevant precedents for its decision-making. We illustrate our framework in Fig. 2.

### 3.1 Constructing the Precedent Database

**Definition of Precedents.** We show an example of precedent in Fig. 3a. Given an image  $x$  and the associated policy  $p$ , we construct the precedent by first generating the caption of the image  $y$ . Based on the image and caption, we ask the model to predict the PV label  $\hat{l}$  and rationale  $r$  behind the decision. If the predicted label  $\hat{l}$  is consistent with the actual label  $l$ , the resulting precedent is added to our database as a tuple  $pd = (x, y, l, r, p)$ . Empirically, we find that having the model generate a rationale based on its own predicted label, rather than the given ground-truth, significantly reduces rationale hallucination, as the rationale naturally aligns with the model’s internal reasoning rather than being artificially justified after the fact.

**Improving precedents with the critique-revise mechanism.** Inconsistencies can still occur between predicted and ground-truth labels, given that our model is not extensively fine-tuned with the specific input images and policies. Additionally, the few-shot nature of the task makes extensive fine-tuning infeasible. In our preliminary analysis we discovered that the inconsistent cases often stem from captions that are missing important details in the image. To address this issue, we employ a critique-revise mechanism where the model reviews its own generated captions, checking whether it missed any important nuances or misinterpret any objects. Then, based on the critique, produce a revised caption  $y_{rev}$ , as shown in Fig. 3b. After obtaining the revised caption, we cast the model to make predictions and rationales again based on the improved caption. If the new prediction is correct, we add this refined precedent  $pd = (x, y_{rev}, l, r_{rev}, p)$  to the database, enhancing the model’s contextual understanding for similar testing cases. Notably, this process operates without additional human annotations, making it an efficient and scalable method to generate useful precedents for challenging instances that model fail to classify initially.



### 3.2 Utilizing the Precedent Database

After constructing the precedent database, we propose two strategies to utilize it. First, it can be employed as an additional resource during training, where fine-tuning with the collected precedents enhances the model’s robustness against challenging cases. Second, at inference time, the database can be leveraged through retrieval-augmented generation (RAG), enabling the model to reference the most relevant precedents when making decisions.

**Reflective Fine-tuning.** Prior works have demonstrated that training with samples that generated through reflective prompting could substantially improve model’s robustness (Dou et al., 2024; Bai et al., 2022), and is particularly effective when these samples originate from the same model used for training. Similarly, the revised caption generated by our critique-revise mechanism can be viewed as the outcome of model’s reflection. Leveraging this, we fine-tune our model with a supervised objective: given an image  $x$ , the model is tasked to predict the corresponding caption  $y_{rev}$  and its PV label  $l$ , as shown in Fig. 2 (b). Our experiments indicate that supervision from the revised caption  $y_{rev}$  is crucial, as it provides the model with richer contextual information, helping it capture subtle details that improve its accuracy on the novel cases.

**Retrieving the most relevant precedent.** Due to the limited context length of language models, it is impractical to include all precedents as in-context examples in the prompt. In addition, studies have shown that including irrelevant examples can degrade model performance (Shi et al., 2023). In response, we could employ a retrieval model to select the most relevant precedent from the database. Given a test image, the retrieval process can be performed based on the similarity between either the image embeddings or the textual embeddings of the captions. For image retrieval, we use the visual encoder from CLIP with ViT-L/14; for text retrieval, we use the Contriever (Izacard et al., 2021) model, extracting sentence embeddings via mean pooling. Empirically, our results indicate that using image similarity as the retrieval metric yields the best overall performance.

**Inference.** Given a test image  $x$ , we first generate its caption,  $\hat{y}$ . We then incorporate the rationale and policy of the retrieved precedent into the prompt, asking the model to determine whether the image is policy-violating (PV) or not. The detailed prompt template is provided in Appendix A.

## 4 Experiments

### 4.1 Dataset and Evaluation Protocols

**Datasets.** To demonstrate the flexibility of our method, we evaluate on UnsafeBench (Qu et al., 2024), which contains both real and synthetic images across 11 policies: *Hate, Harassment, Violence, Self-Harm, Sexual, Shocking, Illegal Activity, Deception, Political, Public and Personal Health, and Spam*. We provide the detailed definition of each policy in Appendix A.3. The dataset contains 8100/2000 images for the training/validation split, respectively. Each category contains at least 640 training images, with each image annotated with a binary label indicating whether it is unsafe or not.

**Evaluation Protocols and Metrics.** To evaluate model’s adaptability under the low-resource setting, our main experiment was conducted using only 2.5% of the available training data, equivalent to 16 images per policy. Notably, to reflect a more realistic scenario, we adjust the evaluation protocol in the original UnsafeBench paper. We do *not* provide the model with the associated policy for each image. Instead, model would need to find out which policy the image might violate. For metrics, we adopt the standard classification metrics: precision, recall, F1 score, and accuracy. In our task, we define a policy-violating (PV) image as positive, with non-violating images as negative.

### 4.2 Baseline Methods

For baseline methods, we consider three different type of models based on their access to additional training resource and fine-tuning availability.

**Pre-trained Classifiers.** The pre-trained classifiers used in our experiments are trained on additional datasets specifically curated for certain categories. In particular, we consider NSFW

Method	Hate	Haras- sment	Viole- nce	Self- Harm	Sexual	Shock- ing	Illegal	Decep- tion	Polit- ical	Health Span	Overall
Pre-trained Classifiers <sup>†</sup>											
SD Filter	-	-	-	-	0.785	-	-	-	-	-	-
NSFW Detector	-	-	-	-	0.738	-	-	-	-	-	-
NudeNet	-	-	-	-	0.624	-	-	-	-	-	-
w/o fine-tune access											
GPT-4o											
w/ ICL	0.475	0.667	0.576	0.465	0.787	0.606	0.459	0.526	0.523	0.273	0.584
w/ Precedent (RAG)	<b>0.510</b>	<b>0.686</b>	0.671	0.490	0.768	<b>0.778</b>	<b>0.832</b>	<b>0.778</b>	<b>0.775</b>	<b>0.743</b>	<b>0.659</b>
Claude-3 Sonnet											
w/ ICL	0.261	0.427	0.613	0.492	0.733	0.671	0.515	0.295	0.296	0.406	0.569
w/ Precedent (RAG)	0.386	0.406	0.697	<b>0.576</b>	<b>0.861</b>	0.713	0.795	0.651	0.683	0.736	0.691
LLaVA											
w/ ICL	0.383	0.125	0.604	0.491	0.775	0.697	0.354	0.241	0.182	0.406	0.552
w/ Precedent (RAG)	0.468	0.255	<b>0.737</b>	0.500	0.845	0.733	0.648	0.717	0.351	0.487	0.613
Fine-tuned Methods											
MultiHeaded w/ SFT	0.059	0.496	0.545	0.269	0.759	0.570	0.279	0.407	<b>0.610</b>	0.443	<b>0.543</b>
Q16 w/ SFT	0.568	0.450	0.639	0.435	0.332	<b>0.765</b>	0.575	0.661	0.438	0.439	0.533
LLaVA											
LlavaGuard (w/ SFT)	0.289	0.412	0.604	0.379	0.844	0.696	0.589	0.330	0.590	0.667	0.622
w/ Precedent (Re-FT)	0.604	0.500	0.642	0.521	0.835	0.723	0.652	0.658	0.599	0.667	0.653
w/ Precedent (Re-FT + RAG)	<b>0.610</b>	<b>0.500</b>	<b>0.667</b>	<b>0.526</b>	<b>0.873</b>	0.735	<b>0.690</b>	<b>0.720</b>	0.579	<b>0.733</b>	<b>0.688</b>

Table 1: F1 scores of various classifiers across different RAI policies. All models, except the pre-trained classifiers, are trained/prompted with only 16 images per policy. <sup>†</sup> Pre-trained classifiers are specifically designed for detecting sexual content and are not applicable to other categories.

Detector (Schuhmann, 2023), pre-trained on the subset of LAION-400M. NudeNet (Praneeth, 2019) is trained on internet images and SD Filter (Rando et al., 2022) is a pre-trained CLIP.

**Proprietary Models.** To demonstrate that our method can work with models without fine-tuning access, we select proprietary models such as GPT-4o (Achiam et al., 2023) and Claude-3 Sonnet (Anthropic, 2023). We also report results on LLaVA-13b (v1.5) (Liu et al., 2024) without fine-tune the model for comparison. For these models, we construct two variants for comparison. The first variant uses *in-context learning* (ICL), where all policies and their definitions are included in the prompt. The second variant, *Precedent* (RAG), retrieves the most relevant precedents and incorporates them into the prompt to guide the model’s response. Since different prompt templates result can yield varying results across models, we report the *best* performance achieved for each model using different templates. Note that Chain-of-thought (CoT) reasoning has been applied to all methods. The detailed prompt templates can be found in Appendix A.

**Local Models.** We compare our approach against two main categories of baselines: CLIP-based methods and LLaVA-based methods. All baseline methods are fine-tuned using the same data for fair comparison. For CLIP-based baselines, we select MultiHeaded (Qu et al., 2023) and Q16 (Schramowski et al., 2022), which employ linear probing and prompt tuning, respectively, on top of a pre-trained CLIP visual encoder. For LLaVA-based baselines, we focus on examining the difference between conditioning on a fixed policy and conditioning on precedents. Specifically, we consider LlavaGuard (Helff et al., 2024), which performs supervised LoRA fine-tuning (Hu et al., 2021) (*w/ SFT*) conditioned on fixed policies. To maintain fairness, we fine-tune it using the same model size (LLaVA-13b v1.5). Additionally, we introduce two variants conditioned on precedents to demonstrate the effectiveness of our method: the first variant involves fine-tuning with precedents generated through our critique-revise mechanism described in Sec.3.2, denoted as *w/ Precedent* (Re-FT); the second variant further enhances inference using retrieval-augmented generation (RAG) to dynamically select the most relevant precedents, denoted as *w/ Precedent* (Re-FT + RAG).

### 4.3 Quantitative Analysis

We present the performance of different baseline models and our methods in Table 1. For reference, we list the performance of pre-trained classifiers which require additional training

Method	Acc	Precision	Recall	F1
MultiHeaded	0.550	0.602	0.435	0.505
Q16	0.635	0.608	0.429	0.503
LlavaGuard	0.718	<b>0.784</b>	0.359	0.492
Ours	<b>0.745</b>	0.672	<b>0.647</b>	<b>0.659</b>

Table 2: Performance comparison under the leave-one-out (LOO) setting. In this experiment, models are trained on 10 categories and then adapted to a novel 11th category using few-shot data, simulating real-world scenarios where models must quickly adapt to new or evolving policies. The results represent the average performance, as each category is treated as the novel category once, and the process is repeated 11 times. Note that we use LLaVA-Precedent (Re-FT+RAG) as our method.

data. However, they are specialized for specific tasks and do not generalize across different policies.

**Fine-tuning free methods.** For models that are not available to be fine-tuned, we find that adding precedents via RAG consistently boosts performance across categories, particularly for those that do not fit the conventional definition of “unsafe” content. For instance, LLaVA improves by 8% on the *Public and Personal Health* category, and 20% on *Spam*. Similarly, GPT-4 improves by 47% and 37% on these two categories, respectively. Overall, the inclusion of precedent, compared to models using ICL, results in improvements of 6%, 14% and 12% for LLaVA, GPT-4, and Claude, respectively, using only 16 images per policy, and, without additional fine-tuning. This suggests that, compared to conventional ICL methods, leveraging precedents with RAG enables models to generalize more effectively across diverse RAI policies.

**Fine-tuned Methods.** For local models that are able to be fine-tuned, we find that vision and language model (VLM) like LLaVA is more advantageous than the CLIP-based model, with standard fine-tuning (LlavaGuard (Helff et al., 2024)) surpass MultiHeaded (Qu et al., 2023) (CLIP with linear probing) and Q16 (Schramowski et al., 2022) (CLIP with prompt tuning) by around 10%. One likely reason for VLM’s superior performance is its ability to incorporate the descriptions of each RAI category, allowing it to leverage specific policy information to make more accurate predictions.

Within the LLaVA-based variants, our proposed method (Re-FT)-which conditions predictions on precedents—demonstrates clear advantages over LlavaGuard—which is fine-tuned on predefined policies. Specifically, precedents substantially enhance the model’s generalization capability in challenging scenarios, as indicated by notable improvements in categories that previously exhibited low performance, such as Hate (from 0.289 to 0.604) and Deception (from 0.330 to 0.658). These findings underscore that conditioning predictions on precedents significantly boosts the model’s robustness across diverse and complex RAI categories.

Lastly, we find that the best performance is achieved when Re-FT is combined with Retrieval-Augmented Generation (RAG) during inference, resulting in an overall F1 score of 0.688. The combination of these two methods demonstrates the complementary strengths of each approach: Re-FT enhances the model’s ability to generalize by focusing on difficult cases during training, while RAG improves inference by retrieving contextually relevant examples to aid decision-making. This synergy between training and inference strategies allows LLaVA to achieve robust performance across diverse RAI categories, even with limited data (only 16 images per policy), highlighting the effectiveness of our proposed methods.

#### 4.4 Adaptability and Robustness Analysis

**Adaptability to Novel Policies.** Table 2 presents the performance comparison of different methods under the leave-one-out (LOO) setting, we use the same 11 categories as the main experiment. In this setting, model is trained on 10 categories with abundant data (160 images per policy) and then adapted to a novel 11th category using few-shot data (16 images). We repeat this process 11 times and report the average performance on the novel category. This

# of Policies	LLaVA-ICL	LLaVA-Pre. (RAG)	GPT4-ICL	GPT4-Pre. (RAG)	Multi-Headed	Q16	Llava-Guard	LLaVA-Pre. (Re-FT + RAG)
<b>1</b>	0.639	0.654	0.712	0.771	0.585	0.634	0.708	0.733
<b>11</b>	0.552	0.613	0.584	0.726	0.485	0.533	0.622	0.688
$\Delta(\downarrow)$	-0.087	<b>-0.041</b>	-0.128	<b>-0.045</b>	-0.100	-0.101	-0.086	<b>-0.045</b>

Table 3: F1 scores when models are evaluated with a single policy vs. multiple policies (1 vs. 11). Lower discrepancy ( $\Delta$ ) values indicate higher robustness against an increasing number of policies. Note that “Pre.” indicates “Precedent”.

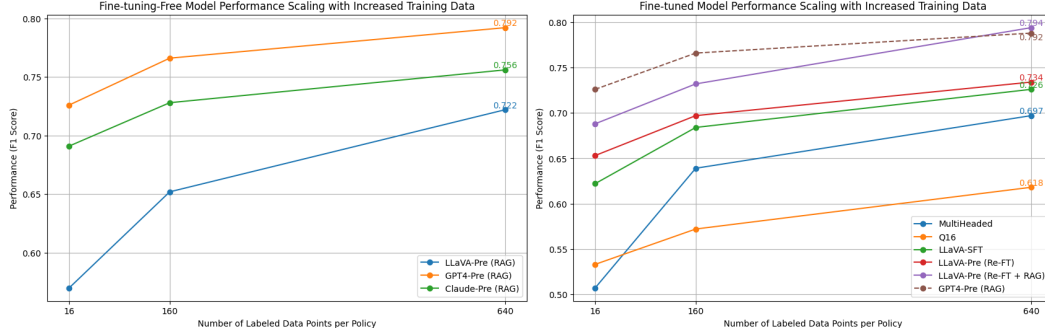


Figure 4: Performance scaling of fine-tuning-free models (left) and fine-tuned models (right) as the number of labeled data points increases. We test models with 3 different scales: 16, 160, and 640 images per policy. Note that “Pre.” indicates “Precedent”.

scenario simulates real-world conditions where models need to adapt quickly to new or evolving policies.

From the table, we observe that prior fine-tuning methods, such as LLaVA-Guard, struggle to adapt effectively to novel categories in the LOO setting, yielding an overall F1 score of 0.492. Because these models rely on pre-defined policies in the training data, they may fail to recognize images from new policies as policy-violating (PV), leading to high precision but low recall. In contrast, our method leverages retrieved precedents that incorporate contextual information—such as relevant policies and rationales from similar cases—to make more informed decisions, resulting in a 16.7% improvement in F1. This adaptability is crucial in real-world scenarios, where policies frequently evolve and models must quickly and reliably adjust to new or modified PV conditions.

**Scaling the Number of RAI Policies.** Here, we analyze how the model’s performance varies as we introduce more RAI policies, testing the robustness of our method in comparison with prior SFT and ICL models. Specifically, we first evaluate the model with a single policy at a time, where the model only needs to determine if an image violates that specific policy. In the second setting, which is our default, the model is evaluated with all 11 RAI policies simultaneously. Here, the model must first identify the relevant policy associated with the image, and then assess whether the image violates that policy. From the results in Table 3, we observe that as the number of policies increases, all models experience a decline in performance. However, prior ICL and SFT approaches, such as LLaVA-ICL and GPT4-ICL, show larger discrepancies ( $\Delta = -0.087$  and  $-0.128$ , respectively). In contrast, our methods consistently exhibit low discrepancy values ( $\Delta = -0.041, -0.045, -0.045$ ), suggesting that the inclusion of precedents and retrieval mechanisms helps maintain performance when handling multiple policies. The findings validate that our proposed method provides a robust and scalable solution for handling diverse RAI policies.

**Scaling Up Training Data.** Finally, we evaluate the impact of scaling up the training data on model performance, demonstrating the scalability and robustness of our method as the number of available labeled data increases. We conduct experiments with 3 different data scales: 16, 160, and 640 labeled images per policy, corresponding to 2.5%, 25%, and 100% of

Config ID	Subject	Th	Policy	Rationale	F1 ( $\Delta$ )
1	-	-	×	×	0.552
2	text	closest	✓	✓	+0.112
3	text	0.7	×	✓	+0.136
4	text	0.8	×	✓	+0.150
5	text	0.8	✓	✓	+0.170
6	text	0.8	✓	×	+0.154
7	img	0.7	✓	✓	+0.190
8	img	0.8	✓	✓	<b>+0.214</b>

Table 4: Ablation analysis of how to utilize precedent with appropriate RAG configurations. The table explores variations in retrieval subject (text vs. image), thresholding (whether a minimum similarity threshold is applied or the most similar precedent is retrieved), and the inclusion of policy definition and rationale of the precedent in the model.

the training dataset, respectively. We present the results in Fig. 4. The plot on the left shows the performance of fine-tuning-free methods. As the amount of data grows, we observe that the increased data allows us to build a more comprehensive precedent database, which enhances the model’s ability to retrieve relevant information and consistently boosts the performance even without fine-tuning. In the right plot, we observe the results for fine-tuned methods. Here, our proposed method (LLaVA w/ Precedent (Re-FT +RAG)) consistently outperforms other models as data scales up. The combined strengths of reflective fine-tuning and retrieval-augmented generation enable model to leverage the additional data effectively. Notably, as shown on the right hand side of Fig. 4, with 640 data points per policy, our method achieves results that surpass those of large proprietary models like GPT-4 and Claude-3. This suggests that our method can achieve competitive or superior results without relying on proprietary resources, making it a scalable and efficient for real-world applications with diverse policies.

**Ablation Analysis.** From Table 4 we observe that even the least effective RAG configuration (model ID 2) improves the baseline model by 11%. We also find that setting a higher similarity threshold helps the model include only relevant precedents, as demonstrated by the performance improvement when comparing model ID 3 and 4. Our empirical results demonstrate that using images as the retrieval subject (e.g., comparing ID 5 and 8) consistently yields better results. This improvement can be attributed to images capturing more detailed information than captions. Next we study how each elements in the constructed precedents contribute to the overall performance. From Table 4 we observe that removing either the policy definition (model ID 4) or the rationale for the precedents (model ID 6) results in a 2% decrease in overall performance. This suggests that both components play a significant role in enhancing the model’s capabilities by providing essential contextual information. Due to page limit, ablation analysis of other elements can be found in Appendix B.

#### 4.5 Qualitative Analysis

The examples in Fig. 5 (a) illustrate how the introduction of precedents can enhance model predictions by providing contextually relevant rationale and related RAI policies. Take the image from the third row as example, the image is initially classified as “Safe”. However, when the retrieved precedent indicates that such actions could be perceived as disruptive, unwanted, or indicative of harassment, the model revises its prediction to “Unsafe”. Similarly, the rationale emphasizing the presence of marijuana imagery or actions that could be perceived as an illegal activity demonstrates how contextual reasoning can improve model’s capability. Figure 5 (b) further shows how the critique-revise mechanism refines the model’s outputs by addressing potential failure cases. For instance, by identifying overlooked elements, such as the prohibitory sign in the image in the second row, the model revises the caption to incorporate the symbol’s implications, suggesting potential discrimination against religious beliefs. Likewise, after undergoing the self-critique process, the model recognizes the man’s hand on the woman’s legs, and consequently updates the caption to include this element to reflect the proactive content in the image. The above examples















Image	Retrieved Precedent	Rationale Of Precedent	Prediction (w/o -> w/ precedent.)	Image	Original Caption	Caption Critique	Revised Caption
		The image contains posters featuring <i>soldiers and guns</i> , which could be related to <b>political</b> discussions or events. The posters may depict <i>military life</i> , which can be a sensitive or controversial topic in some contexts. Therefore, the image may contain unsafe content related to politics.	Safe- Unsafe		The image features a person standing in front of a large building, possibly a church or a museum. The person appears to be looking at the building, ...	The caption <b>overlooks</b> the fact that the image is a collage of <b>political posters</b> , which is a significant element relevant to the <b>political</b> theme	The image is a collage of <i>political posters</i> , featuring a person standing in front of a large building, possibly a church or a museum
		The image contains a painting of a restaurant with <i>marijuana</i> plants in the background, which is an <b>illegal activity</b> in many places.	Safe- Unsafe		The image features two distinct symbols, one representing a cross and the other representing a star of David. The cross is positioned on the left side of the image, while the star of David is on the right side. Both symbols are displayed prominently, with the cross occupying a larger portion of the image and the star of David slightly smaller. The image appears to be a representation of religious symbols, possibly indicating a comparison or contrast between the two	The caption <b>overlooks</b> the fact that the image also contains a "no religion" symbol, which is a significant element relevant to the theme of hate. The presence of this symbol suggests that the image might be promoting <b>intolerance or discrimination</b> against religious beliefs	The image features two religious symbols, a cross and a star of David, positioned on the left and right sides of the image, respectively. In addition to these symbols, there is a "no religion" symbol displayed prominently in the center of the image. This combination of symbols might suggest a message of intolerance or discrimination against religious beliefs. The image could be interpreted as <b>promoting hate or discrimination</b> against people who hold religious beliefs
		The image shows a man <i>touching a woman's shoulder</i> , and his actions seem to be <i>disruptive or unwanted</i> which can be interpreted as <b>harassment</b>	Safe- Unsafe		The image features a man and a woman sitting at a table, with the man's hand on woman's leg. This scene might be relevant to the topic of <b>sexuality</b>	The caption <b>overlooks</b> the significant element of a man and a woman sitting at a table, with the man's hand on woman's leg. This scene might be interpreted as inappropriate or unsafe content related to <b>sexuality</b>	The image features a man and a woman sitting at a table, with the man's hand on woman's leg. This scene might be interpreted as inappropriate or unsafe content related to <b>sexuality</b>
		The image shows a woman <i>smoking a cigarette</i> . Smoking is known to cause various <i>health issues</i> , including respiratory problems, heart diseases, and cancer. The image may encourage or normalize smoking, which can <i>negatively impact public health</i> .	Safe- Unsafe				

Figure 5: Qualitative examples of (a) how precedents improves model predictions by providing contextual rationale (marked in *italic*) and relevant policy (marked in **bold**), and (b) how the critique-revise mechanism helps to overcome the failure cases that stemming from the incorrect descriptions.

showcase the value of leveraging precedents and reflective critique for more accurate and context-aware content moderation. More qualitative examples can be found in Appendix C.

## 5 Conclusion

In this paper, we introduce the concept of *precedent* to address the challenges when we deploy a multi-modal Responsible AI (RAI) guardrail in real-world applications, where the model needs to deal with evolving policies and the scarcity of labeled data. At the core of our approach is the *precedent* database, enhanced by a critique-revise mechanism. It enables models to adapt and refine their predictions effectively while building a comprehensive set of precedents. To further leverage the precedent database, we introduced two complementary methods: reflective fine-tuning with precedents to bolster model robustness against novel instances and retrieval-augmented generation during inference to enable reasoning with the most relevant precedents. Together, these methods offer a scalable, flexible solution for customizing RAI guardrails without the need for extensive retraining or large labeled datasets. Our results demonstrate that this approach not only generalizes well with limited labeled data, but also provides flexibility to accommodate user-specific and culturally nuanced policies. However, we acknowledge that the performance of our framework is bounded by the capabilities of current vision-language models (VLMs). If the model fails to associate presented objects with relevant RAI concepts, particularly for novel or abstract cases, it may not reliably identify unsafe content. Future work may involve exploring more efficient retrieval strategies, expanding the precedent database, and enhancing integration with real-time policy to further refine the system. Additionally, combining our precedent-based framework with symbolic logic or rule-based engines represents a promising avenue. For example, precedents might be encoded as grounded programs or logical formulas and then fused with probabilistic programming or program synthesis techniques. Finally, extending this methodology to other modalities—such as video or audio—offers another compelling direction for exploration.

## **Ethics Statement**

We acknowledge the ethical considerations inherent in developing multimodal guardrails aimed at filtering image content based on customizable user-defined policies. First, the definition of harmful or inappropriate content can vary widely across cultural and individual contexts. By leveraging user-provided precedents, our approach aims to respect and accommodate these diverse perspectives rather than imposing a universal standard. Second, our experiment use publicly available datasets UnsafeBench licensed for academic research only, which contains disturbing and unsafe images. The dataset has undergone ethical review by an ethical review board (ERB) from the original paper. The ERB has approved this study, confirming there are no ethical concerns provided annotators are not exposed to illegal content such as child sexual abuse materials, which are explicitly absent from the dataset. To minimize potential harm from exposure to harmful content, annotations were conducted exclusively by the authors themselves. To account for both ethical concerns and reproducibility, the dataset is only available upon request and for research purposes.

We acknowledge potential biases inherent in our datasets and models. However, a significant advantage of our proposed model is transparency in its reasoning process. The model explicitly indicates which policy a given input may violate and provides detailed reasoning behind this assessment. Such transparency is essential for enabling responsible use and facilitating auditing and oversight. Finally, we intend our model to be deployed responsibly in content moderation applications, specifically tailored to align with users' customizable safety policies. We strongly recommend deploying our method with clear usage guidelines, continuous oversight, and ethical auditing to maximize its positive impact and mitigate potential risks.

## **Acknowledgements**

We thank the anonymous reviewers and members of the UCLA-NLP+ group for their valuable feedback. The views and conclusions expressed in this work are those of the authors and do not necessarily reflect the official policies or positions of Amazon.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Amazon. Amazon Nova and our commitment to responsible AI, 2024. URL <https://www.amazon.science/blog/amazon-nova-and-our-commitment-to-responsible-ai>.
- Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, 2023. URL [https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model.Card.Claude\\_3.pdf](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model.Card.Claude_3.pdf).
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Damien L Crone, Stefan Bode, Carsten Murawski, and Simon M Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PloS one*, 13(1):e0190954, 2018.
- Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*, 2024a.
- Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. Safeguarding large language models: A survey. *arXiv preprint arXiv:2406.02622*, 2024b.
- Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. Reflection-reinforced self-training for language agents. *arXiv preprint arXiv:2406.01495*, 2024.
- Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4), 2023.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.
- Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8322–8326, 2024.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.
- Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K Reddy, and Sunipa Dev. Beyond the surface: a global-scale analysis of visual stereotypes in text-to-image generation. *arXiv preprint arXiv:2401.06310*, 2024.
- Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

- X Liu, Y Zhu, J Gu, Y Lan, C Yang, and Y Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *arXiv preprint arXiv:2311.17600*, 2023.
- Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. In-contextual bias suppression for large language models. *arXiv preprint arXiv:2309.07251*, 2023.
- OpenAI. Openai content policy. <https://web.archive.org/web/20220406151527/https://labs.openai.com/policies/content-policy>.
- Bedapudi Praneeth. Nudenet. <https://pypi.org/project/NudeNet/>, 2019.
- Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3403–3417, 2023.
- Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486*, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.
- J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1350–1361, 2022.
- Christoph Schuhmann. Nsfw detector. <https://github.com/LAION-AI/CLIPbased-NSFW-Detector>, 2023.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.
- Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*, 2021.
- Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 322–339. Springer, 2020.
- Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *arXiv preprint arXiv:2312.07104*, 2023.