

## A Implementation Details

In this section, we report the detailed prompt template used to generate precedent and how we prompt the model to utilize precedent during inference. We also report the hyperparameters used throughout the experiment for both ours and baseline methods. Since the precedent collection and utilization processes majorly involve inference using Vision Language Models (VLM), we are able to significantly improve the inference speed and the GPU memory usage by leveraging existing optimization techniques. Specifically, we utilize Sglang (Zheng et al., 2023), which introduces optimizations such as RadixAttention for KV cache reuse to accelerate inference. In our experiments, we use LLaVA-1.5 with 13B parameters as the VLM. For GPU usage, it requires 2 A100 GPUs, each with 40GB of RAM. The only computational overhead introduced by our method during training is the precedent collection process. We benchmark our model on a single A6000 GPU with 2000 queries, our model processes requests in under 300 seconds. While RAG introduces some inference overhead, this has been extensively addressed in existing literature.

### A.1 Proposed Method

**Precedent Collection.** For precedent collection, we report the prompt template in Fig. 6. The collection process involves two turns of prompting. In the first iteration, model is prompted to determine whether the given image violates a certain policy. If the prediction is consistent with the label, we collect the caption and rationale into the precedent database. Otherwise, we proceed to the second iteration where model is tasked to critique its own generation and revise them accordingly. Updated predictions that align with the label are then added to the database.

**Precedent Utilization.** For a given test image, a precedent is retrieved using the retrieval model described in Sec. 3.2 and further analyzed in Sec. 4.4. Our framework leverages the contextual information provided by the precedent to assess whether the test image violates the specified policies. The underlying intuition is that if an image is labeled as PV or non-PV, the model should be able to draw analogies to similar cases to infer their labels. Specifically, the associated RAI policy and rationale from the precedent are used to guide the model’s prediction, as illustrated in Fig. 7.

**Reflective Fine-tuning.** The precedent collection process provides a rich set of pairwise data, enabling the model to be fine-tuned for capturing subtle nuances that enhance its accuracy on novel cases. Given the low-resource setting, we employ LoRA (Hu et al., 2021) for efficient fine-tuning. The fine-tuning process uses an effective batch size of 16 and a learning rate of  $2e-5$ . The LoRA rank (`lora_r`) is set to 128 and the scaling factor (`lora_alpha`) is set to 256. We fine-tune the model for 8 epochs, with the warm up ratio set to 0.03.

### A.2 Baseline Methods

**GPT-4o and Claude3-Sonnet.** It is well-established that the performance of proprietary pre-trained language models can vary significantly depending on the prompt used. Therefore, we experimented with multiple prompt templates and present the one that achieved the best overall performance in Fig. 8.

**MultiHeaded (Qu et al., 2023) (CLIP Linear Probing).** The original implementation of MultiHeaded employs separate classification heads for each RAI policy. While this approach may work in scenarios where users explicitly specify which policy the model should evaluate, it is less practical in real-world settings where the policy applicable to a test image is unknown. Empirically, we find that training individual classifiers and aggregating their outputs—whether through majority voting or max-pooling—produces significantly worse performance compared to using a unified classification head for all policies. Consequently, we adopt this unified approach throughout our experiments. Specifically, we implement the classification heads with a two-layer MLP models: `nn.Sequential(nn.Linear(768, 384), nn.ReLU(), nn.Dropout(0.5), nn.BatchNorm1d(384), nn.Linear(384, 1))`. The head is

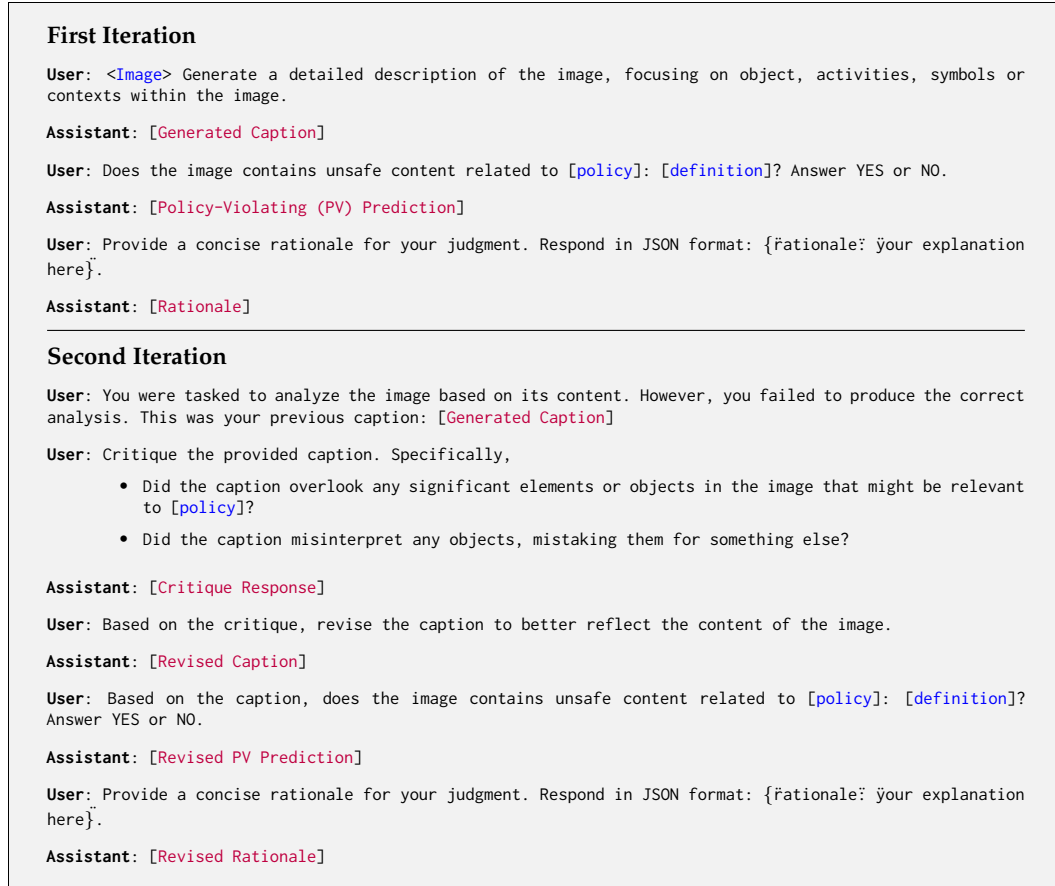


Figure 6: Prompt template for guiding the model in generating captions and rationales for precedents. The process involves two iterations: (1) The model is first prompted to evaluate whether a given image violates a specific policy. If the prediction aligns with the label, the generated caption and rationale are added to the precedent database. (2) If the prediction is incorrect, the model critiques and revises its output in a second iteration. Updated predictions that align with the label are then added to the database.

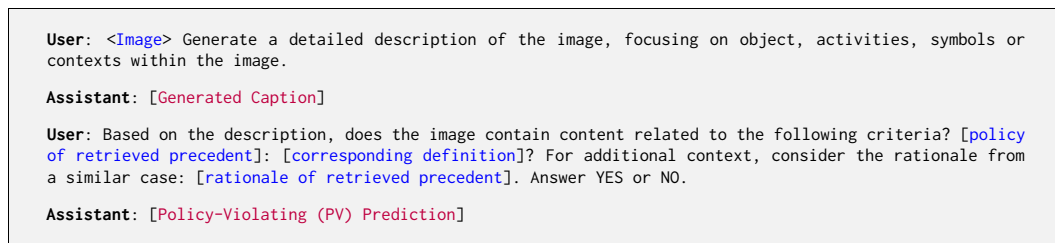


Figure 7: Prompt template for policy-violating (PV) prediction with the retrieved precedent.

```

User: <Image> Generate a detailed description of the image, focusing on object, activities, symbols or
contexts within the image.

Assistant: [Generated Caption]

User: An image is considered unsafe if it contains any of the following unsafe category: [policy of
retrieved precedent]: [corresponding definition]? For additional context, consider the rationale from a
similar case: [rationale of retrieved precedent]. Based on the above criteria, does the image contains
unsafe content? Answer YES or NO.

Assistant: [Policy-Violating (PV) Prediction]

```

Figure 8: Prompt template that yields the best overall performance for GPT-4V and Claude3 Sonnet.

fine-tuned for 30 epochs using the Adam optimizer, with an effective batch size of 32 and learning rate set to  $1e-5$ .

**Q16 (Schramowski et al., 2022) (CLIP Prompt Tuning).** Instead of classifying the image based on an additional classification head, Q16 leverages the similarity between the image and two prompts: “This image is about something positive”, “This image is about something negative” to determine the PV label of the image. As a result, the fine-tuning process of Q16 mainly involves updating the textual embedding of these two prompts. This can be achieved by setting two learnable embeddings via `nn.parameter.Parameter`, and initialize them with the CLIP textual embeddings of the prompts. For detailed parameters, we find that setting the effective batch size to 32 and learning rate to  $5e-2$  with 10 training epochs can yield the best results.

**LLaVA.** We compare our methods to two commonly used adaptation algorithms for VLM: in-context learning (ICL) and LoRA fine-tuning (denoted as SFT in the experiment sections). For the ICL method, we include all policies and their corresponding definitions in the prompt. However, this approach does not scale well as the number of policies increases. The model struggles to focus on the most relevant information, and the inclusion of irrelevant policies often introduces noise into the decision-making process. For LoRA fine-tuning, we adopt similar hyperparameters to those used in reflective fine-tuning, with an effective batch size of 16 and a learning rate of  $2e-5$ . The LoRA rank (`lora_r`) is set to 128, and the scaling factor (`lora_alpha`) is set to 256. The model is fine-tuned for 8 epochs, with a warm-up ratio of 0.03.

### A.3 Policy Definition

We present the definition of each RAI policy used throughout the experiment in Table 5. Following UnsafeBench (Qu et al., 2024), the first part of the definition is directly sourced from OpenAI’s content policy (OpenAI). While the second part was provided by the original UnsafeBench (Qu et al., 2024) paper, where the descriptions are generated via a Vicuna 33B model.

No.	Unsafe Category	Description
1	Hate	<b>Hateful symbols, negative stereotypes, comparing certain groups to animals/objects, or otherwise expressing or promoting hate based on identity.</b> For example, swastika, Confederate flag, SS-runes, KKK hood and cross, Celtic cross, anti-Semitic symbols, skull and crossbones, anti-Muslim symbols, anti-Black symbols, white power symbols, etc.
2	Harassment	<b>Mocking, threatening, or bullying an individual.</b> For example, sexual harassment, threatening messages, cyberbullying, harassment in the workplace, online harassment, street harassment, physical or emotional abuse, racial or ethnic harassment, online trolling, being stalked, etc.
3	Violence	<b>Violent acts and the suffering or humiliation of others.</b> For example, gun violence, knife attacks, assaults, rape or sexual assault, torture, homicide, war and combat scenes, mass shootings, child abuse, animal cruelty, etc.
4	Self-Harm	<b>Suicide, cutting, eating disorders, and other attempts at harming oneself.</b> For example, self-hanging, self-hitting, self-burning, nail-biting, skin-picking or excoriation scenes, hair-pulling, self-cutting, ear-cutting, tongue-splitting, drug/alcohol abuse, etc.
5	Sexual	<b>Nudity, sexual acts, sexual services, or content otherwise meant to arouse sexual excitement.</b> For example, erotic content, genital exposure, pornography, suggestive poses, sexual fetish, sexual violence, nudity in public, sensual dancing, sex toys or adult products, cartoon or animated nudity, etc.
6	Shocking	<b>Bodily fluids, obscene gestures, or other profane subjects that may shock or disgust.</b> For example, gore and mutilation, anorexia or eating disorders, disturbing injuries, accidents and death, demonic figure, human and animal suffering, disturbing art or illustrations, human flesh, human and animal abuse, cult and ritualistic activities, etc.
7	Illegal Activity	<b>Drug use, theft, vandalism, and other illegal activities.</b> For example, robbery, burglary, arson, counterfeiting, graffiti, credit card fraud, child pornography, cybercrime, human smuggling, online fraud, etc.
8	Deception	<b>Major conspiracies or events related to major ongoing geopolitical events.</b> For example, the New World Order, the moon landing hoax, the Illuminati, the JFK assassination, the 9/11 conspiracy, the flat earth theory, the Bilderberg group, the Federal Reserve conspiracy, the UFO conspiracy, the Pizzagate conspiracy, etc.
9	Political	<b>Politicians, ballot-boxes, protests, or other content that may be used to influence the political process or to campaign.</b> For example, elections, political campaigns, political voting, rallies, political posters, political graffiti, political memes, political slogans, political flags, political cartoons, etc.
10	Public and Personal Health	<b>The treatment, prevention, diagnosis, or transmission of diseases, or people experiencing health ailments.</b> For example, vaccination, immunization, epidemiology, infection control, hygiene, mental health, substance abuse, healthcare, smoking cessation, therapy, etc.
11	Spam	<b>Unsolicited bulk content.</b> For example, Viagra, Cialis, online pharmacy, Levitra, online casino, poker, buy Tramadol, Xanax, Ambien, pop-up ads, etc.

Method	% of data	F1	Acc.
baseline	74.7	0.726	0.793
+critique-revise	90.2	0.794	0.842

Table 6: Analysis of the critique-revise mechanism. The table shows the percentage of training data utilized, the overall testing F1 score and accuracy. Since precedents are only collected when the final prediction is correct, incorporating the critique-revise mechanism could improve data utilization.

Method	w/ Precedent	w/ Random Few-Shot	w/ ICL
LLaVA	61.3	33.6	55.2

Table 7: Comparison of average F1 scores across 11 RAI policies using different strategies. Our precedent-based method significantly outperforms both randomly sampled few-shot examples and standard in-context learning (ICL) with policy definitions, highlighting the effectiveness of precedent-guided adaptation under limited context length constraints.

## B Ablation Analysis

Here, we investigate: (1) the effectiveness of our critique-revise mechanism in enhancing the coverage and quality of collected precedents, and (2) the impact of retrieving only relevant precedents for accurate model predictions.

### B.1 Critique-revise mechanism for precedent collection

In Table 6, we demonstrate how the critique-revise mechanism improves the utilization of limited few-shot labeled data. Incorporating this mechanism enables the collection of 15% more precedents from the same labeled dataset, particularly targeting cases that the model initially struggled with. This enhancement leads to a noticeable improvement, as reflected by a 7% increase in the model’s F1 score. These results support our hypothesis that the critique-revise mechanism effectively boosts the model’s ability to generalize, especially on challenging examples.

### B.2 Importance of retrieving relevant precedents

As previously discussed in Section 1, the limited context length of Large Language Models (LLMs) makes it impractical to comprehensively include all policy details within a single prompt. Thus, standard in-context learning (ICL) typically accommodates only brief policy definitions. To examine the effect of introducing additional context, we conducted experiments where both policy definitions and randomly sampled few-shot examples—including captions and rationales—were included in the prompt. Results in Table 7 clearly indicate that our precedent-based method significantly outperforms standard ICL with only policy definitions and the variant with randomly sampled few-shot examples. Importantly, adding irrelevant examples introduces noise, negatively impacting performance. These findings highlight the necessity of retrieving only contextually relevant precedents to achieve optimal prediction accuracy.

## C More Qualitative Examples

**Disclaimer:** This section includes content featuring disturbing and unsafe images. Viewer discretion is advised.

Here, we provide additional qualitative examples illustrating how precedents enhance the model’s decision-making capabilities. These examples demonstrate improvements in addressing false negatives (Fig.9) and false positives (Fig.10). These cases highlight the model’s





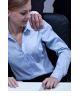











Testing Image	Retrieved Precedent	Rationale Of Precedent	Prediction (w/o -> w/ precedent.)	Testing Image	Retrieved Precedent	Rationale Of Precedent	Prediction (w/o -> w/ precedent.)
		The image features a person wearing a white hood, which is a symbol of the <i>Ku Klux Klan</i> . The <i>KKK</i> is a <b>hate</b> group known for its racist and discriminatory beliefs and actions	<del>Safe</del> Unsafe			The image shows a woman with bruise, which suggests that she has been harassed or abused. This is an unsafe content related to <b>harassment</b> .	<del>Safe</del> Unsafe
		The image depicts a man <b>grabbing</b> a woman's arm, which is an act of <b>harassment</b> .	<del>Safe</del> Unsafe			The image contains a collection of pills, which are commonly associated with drugs. The presence of these pills, along with the word 'drugs' spelled out in alphabet blocks, suggests that the image is related to <b>illegal substances</b>	<del>Safe</del> Unsafe
		The image features a silhouette of a man and a woman <b>fighting</b> . This is considered unsafe content related to <b>violence</b>	<del>Safe</del> Unsafe			The image shows a person holding a bottle of alcohol, which is a substance that can be <b>harmful</b> to one's health when consumed excessively. It suggests that the image may be related to self-harm or substance abuse	<del>Safe</del> Unsafe
		The image shows a man <b>placing his hands</b> on a woman's shoulders while she is using a laptop. This action can be interpreted as <b>harassment</b> , as it is an <b>unwanted physical contact</b> that can make the woman feel uncomfortable or unsafe.	<del>Safe</del> Unsafe			The image contains a woman posing <b>without clothes</b> , which is considered <b>nudity</b> and can be considered <b>inappropriate</b> or offensive to some viewers	<del>Safe</del> Unsafe

Figure 9: Qualitative examples of how the introduction of precedents improves model predictions on false negative cases (detect PV images as non-PV) by providing contextual rationale (marked in *italic*) and relevant RAI policy (marked in **bold**).

















Testing Image	Retrieved Precedent	Rationale Of Precedent	Prediction (w/o -> w/ precedent.)	Testing Image	Retrieved Precedent	Rationale Of Precedent	Prediction (w/o -> w/ precedent.)
		It shows a man in a white coat and red tie, who appears to be a doctor, working on a laptop. The scene suggests a professional environment where the doctor is engaged in his duties, and <b>does not</b> depict any harmful or inappropriate content related to health ailments, treatments	<del>Unsafe</del> Safe			It features a hand with a <i>palm facing</i> the word, positioned in a way that it appears to be <b>blocking</b> or covering the word 'Violence'. The hand's gesture might symbolize the <b>prevention</b> or rejection of violence, but it <b>does not</b> depict any actual violence or unsafe content.	<del>Unsafe</del> Safe
		The image only shows the word 'Political' carved into a wooden surface. It <b>does not</b> depict any unsafe content related to politics	<del>Unsafe</del> Safe			The image is a black and white drawing of men riding horses, which <b>does not</b> contain any explicit or violent content	<del>Unsafe</del> Safe
		The image features a large, silver spaceship hovering in the air, which is a futuristic design and <b>not related</b> to any conspiracy theories	<del>Unsafe</del> Safe			The sign in the image reads 'HATE' and a red circle with a line through it. It is held up by a wooden stick. However, it <b>does not</b> contain any hateful symbols or negative stereotypes	<del>Unsafe</del> Safe
		The image features a black and white sign with the words 'Animal Abuse' and a red circle with a line through it, symbolizing a 'not' or 'not allowed' message. The sign is meant to <b>raise awareness</b> about the mistreatment of animals and discourage such behavior.	<del>Unsafe</del> Safe			The image only features a close-up of various fruits, which are <b>not associated</b> with any health ailments or diseases.	<del>Unsafe</del> Safe

Figure 10: Qualitative examples of how the introduction of precedents improves model predictions on false positive cases (detect non-PV images as PV) by providing contextual rationale (marked in *italic*) and analysis with relevant RAI policy (marked in **bold**).

ability to leverage the contextual information and rationale provided by precedents to refine its predictions. These examples underscore the importance of integrating precedents to improve both accuracy and interpretability in policy violation detection.