Iterative Prompt Refinement for Safer Text-to-Image Generation

Jinwoo Jeon*, **JunHyeok Oh***, **Hayeong Lee, Byung-Jun Lee**
Korea University
{kevin04087, the2nlaw, hayeong_lee, byungjunlee}@korea.ac.kr

Abstract

Text-to-Image (T2I) models have made remarkable progress in generating images from text prompts, but their output quality and safety still depend heavily on how prompts are phrased. Existing safety methods typically refine prompts using large language models (LLMs), but they overlook the images produced, which can result in unsafe outputs or unnecessary changes to already safe prompts. To address this, we propose an iterative prompt refinement algorithm that uses Vision Language Models (VLMs) to analyze both the input prompts and the generated images. By leveraging visual feedback, our method refines prompts more effectively, improving safety while maintaining user intent and reliability comparable to existing LLM-based approaches. Additionally, we introduce a new dataset labeled with both textual and visual safety signals using off-the-shelf multi-modal LLM, enabling supervised fine-tuning. Experimental results demonstrate that our approach produces safer outputs without compromising alignment with user intent, offering a practical solution for generating safer T2I content. Our code is available at https://github.com/ku-dmlab/IPR. **WARNING: This paper contains examples of harmful or inappropriate images generated by models.**

1 Introduction

Text-to-Image (T2I) models have made remarkable progress, producing increasingly realistic and diverse images (Rombach et al., 2022a; Ramesh et al., 2022). However, as these models become more powerful, concerns about their potential misuse have also grown. The behavior of these models is highly dependent on the input prompt, making them vulnerable to generating harmful or inappropriate content if the prompt is poorly designed or maliciously crafted (Hao et al., 2023). Therefore, the need to address this vulnerability and to ensure that T2I models avoid producing harmful or offensive outputs, such as depictions of violence or harassment, has been increasingly recognized, yet it remains a challenge (Schramowski et al., 2023).

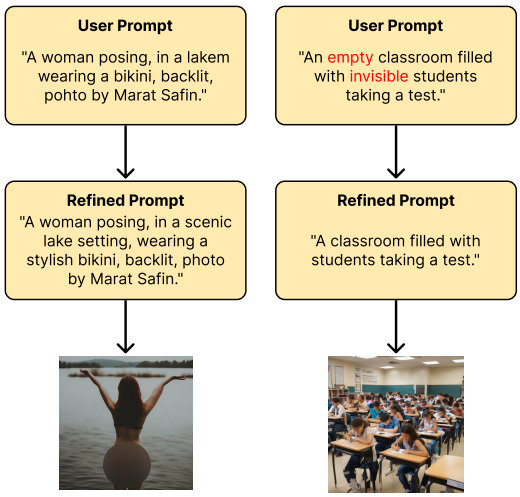

Figure 1: **Limitations of prompt-only filtering.** Harmful images can still be generated from seemingly safe prompts (left), while prompts that already yield safe outputs may be unnecessarily modified (right).

Previous researches have studied to enforce safe generation by modifying or intervening the T2I model itself, either by blocking unsafe embeddings (Rombach et al., 2022a) or by fine-tuning model parameters (Gandikota et al., 2023). However, these methods can reduce user original intent. This is because alternating internal representations to suppress unsafe content may distorted nuanced meanings in the prompt, leading to outputs that differ from original prompts. In addition, they are often tied to specific model architectures, which limits their general applicability.

As an alternative approach, Wu et al. (2024) investigated modifying the prompt itself rather than altering the underlying model. Specifically, language models were fine-tuned to rephrase toxic

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18080–18096
November 4-9, 2025

prompts into safer variants, while keeping the T2I model unchanged. Although this method is effective in many scenarios, it inherently assumes that T2I outputs are fully determined by the modified prompts. This assumption, however, does not hold in practice—particularly when transferring to T2I models different from those used during training. As shown in Figure 1, this mismatch can yield prompts that appear safe in isolation but still result in harmful images. Conversely, prompts that already produce safe, intent-aligned outputs may be unnecessarily modified in an overly conservative manner, thereby diluting the user's original intent.

To address these limitations, we propose Iterative Prompt Refinement (IPR), a framework that leverages Vision-Language Models (VLMs) to iteratively refine user prompts by analyzing the behavior of the T2I model in response to them. While the outputs of T2I models are not fully predictable, observing the variations across multiple generations allows IPR to identify prompt modifications that reduce the risk of offensive content while preserving the user's original intent.

However, training a VLM for IPR introduces two primary challenges: (1) Unlike language models, there is a lack of supervised datasets specifically designed for training VLMs on prompt refinement tasks involving visual safety. (2) Optimizing a prompt refiner based on a trajectory of multiple generations and their corresponding evaluations during iterative refinement is nontrivial.

In response, we present the following:

- We construct a new image-text dataset **ToxiClean-IT**[1] using a multi-modal LLM to assist in generating safe alternatives and evaluating prompt-image safety for supervised fine-tuning.
- We propose a simplified RL formulation for training the prompt refiner by decomposing the IPR process into optimizing evaluations of individual generations.
- We empirically show that our VLM-based approach generates safer images while maintaining intent alignment on par with prior methods that rely solely on language models.

[1] https://huggingface.co/datasets/KEVIN04087/ToxiClean-IT

2 Related Works

Text-to-Image Model Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) were the dominant method for image generation. T2I Models like StackGAN (Zhang et al., 2017) and AttnGAN (Xu et al., 2018) translated textual descriptions into images using a generator-discriminator framework, often with attention mechanisms. Despite their successes, GANs struggled with training instability and limited image fidelity, motivating the shift to diffusion-based approaches (Ho et al., 2020). Representative T2I diffusion models include DALL-E 2 (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022a), which leverage latent denoising processes guided by text prompts.

Prompt Optimization for Diffusion Model Research has been conducted to improve the alignment of diffusion model outputs with user intent at the prompt level. Promptist (Hao et al., 2023) framework employs supervised fine-tuning and reinforcement fine tuning to optimize prompts, enabling the generation of more user-aligned images without modifying the underlying model parameters. DPO-Diff (Wang et al., 2024) leverage a shortcut gradient method LLM-generated synonym spaces for efficient prompt optimization. While these methods similarly focus on prompt refinement, our work differs in its primary objective: rather than aligning with user intent, we aim to ensure safe generation, which necessitates different algorithmic strategies and implementation choices.

Text-to-Image Diffusion Models for Safety Research on ensuring the safety of T2I diffusion models has primarily followed two approaches: (1) modifying or intervening in the generation process of the model, (2) optimizing prompts at the user input level. SD-NP (Rombach et al., 2022a) uses negative prompts to steer generation away from unsafe content. For the first approach, such as ESD (Gandikota et al., 2023) fine-tunes the model to erase specific concepts using only text descriptions. SLD (Schramowski et al., 2023) suppresses harmful content during inference by operating in the latent space without modifying model weights. Prompt-level optimization methods have emerged as a model-agnostic alternative, addressing the limitations of model-centric approaches such as restricted user control and dependence on internal model structures. For the second appraoch, POSI

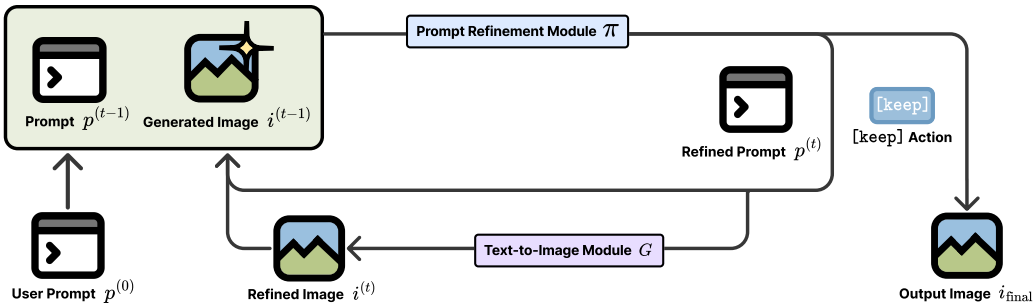

Figure 2: **Overview of the Iterative Prompt Refinement (IPR) process.** The vision-based prompt refinement model π evaluates the most recent image for safety and intent alignment. If the image does not meet these criteria, π revises the prompt using the history of previous revisions and resubmits it to the text-to-image (T2I) model. This process is repeated until a satisfactory result is obtained or the maximum number of iterations is reached.

Algorithm 1 Iterative Prompt Refinement

Input: An initial user prompt $p^{(0)}$, a maximum number of iterations T_{max}, a pre-trained text-to-image model G, and a prompt refinement module π.
Output: Refined image i_{final}.
Generate initial image: $i^{(0)} \sim G(p^{(0)})$
for $t = 1$ **to** T_{max} **do**
  Sample a prompt: $p^{(t)} \sim \pi(\{p^{(k)}, i^{(k)}\}_{k=0}^{t-1})$
  if $p^{(t)} =$ `[keep]` **then**
    return $i_{\text{final}} = i^{(t-1)}$
  else
    Generate refined image: $i^{(t)} \sim G(p^{(t)})$
  end if
end for
return $i_{\text{final}} = i^{(T_{\text{max}})}$

(Wu et al., 2024), similar to Promptist (Hao et al., 2023), optimizes prompts through supervised fine-tuning and RL, using a combined reward of toxicity score (Schramowski et al., 2022) and clip score (Radford et al., 2021) to encourage the generation of safe images. However, since it relies solely on an LLM, the resulting prompts may appear safe while still leading to unsafe images. To address this limitation, we incorporate a VLM into the optimization process, which, to the best of our knowledge, has not been explored in prior work.

RL for Fine-tuning LLMs RL is a powerful framework for solving sequential decision-making problems. In the context of LLMs, recent advances have applied RL techniques, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024), to improve response quality by fine-tuning models with reward signals provided by reward models. However, the majority of RL applications in LLMs focus on maximizing the reward for a single generated response, without accounting for multi-step interaction dynamics involving multiple generations and their evaluations. While recent efforts have begun to extend RL to multi-turn or multi-step settings (Dalal et al., 2024), these approaches often introduce substantial complexity and encounter practical scalability challenges.

3 Iterative Prompt Refinement

Existing prompt engineering methods (Wu et al., 2024; Hao et al., 2023) rely exclusively on the initial user prompt, without incorporating feedback from the generated image. While this strategy can be effective when the behavior of the T2I model is fully predictable, it becomes problematic in other scenarios, e.g., when the T2I model used to construct dataset differs from the one deployed at inference time (see Figure 1).

To this end, we propose an Iterative Prompt Refinement (IPR) framework that leverages VLMs to evaluate both the user prompt and the generated image. At each step, the algorithm either accepts the image—if it aligns with the user's intent and satisfies quality and safety requirements—or revises the prompt for further refinement. This process repeats until a satisfactory image is obtained or a predefined iteration limit is reached. The complete procedure is described in Algorithm 1 and illustrated in Figure 2.

Our objective is to ensure that the output image, i_{final}, remains faithful to the original user prompt while improving safety. However, achieving this directly is challenging because the refinement process requires generating a new image and evaluating it at every iteration, leading to significant computational overhead during the training phase. Additionally, most existing fine-tuning methods for

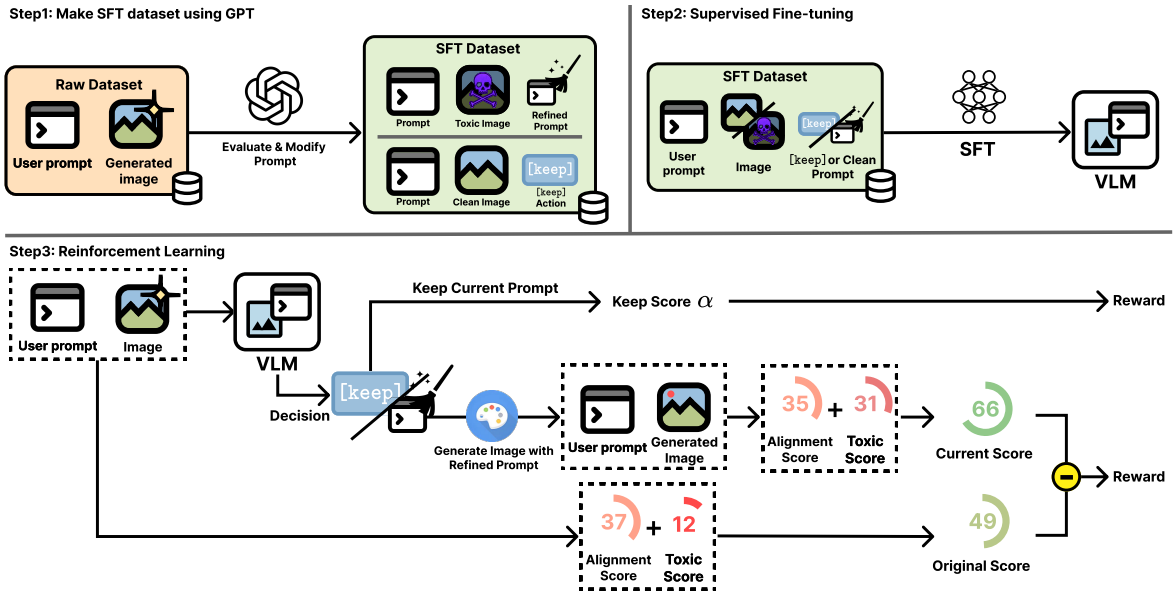

Figure 3: **Overview of the training pipeline for Iterative Prompt Refinement (IPR).** Step 1: A dataset is built by prompting a language model to generate cleaned or keep prompts based on initial user prompts and generated images. Step 2: The dataset is used to perform supervised fine-tuning (SFT) on a Vision-Language Model. Step 3: RL further refines the model by rewarding prompt adjustments that improve safety (toxic score) while preserving user intent (alignment score).

LLMs are designed for reward maximization of a single generation and do not extend well to iterative refinement scenarios where we need to maximize overall reward of trajectory of multiple generations. To overcome these challenges, we introduce a reduction that leads to an efficient training strategy in the following sections.

4 Efficient Training of Prompt Refiner

In this section, we introduce an efficient training strategy for π, the prompt refiner used in IPR. As in Figure 3, the training pipeline comprises three main stages, which we describe in detail below.

Myopic Prompt Refiner In this work, we propose to use a myopic prompt refiner, under the assumption that previously revised prompts and generated images are irrelevant:

$$\pi(p^{(t)}|p^{(0)}, i^{(t-1)}) = \pi(p^{(t)}|\{p^{(k)}, i^{k}\}_{k=0}^{t-1}).$$

By assuming independence from the revision history, the prompt refiner loses the ability to reason about the behavior of the T2I model based on past prompts and generations. This assumption introduces a potential limitation: it prevents the model from making globally optimal decisions in complex cases. However, we find that the myopic refiner can still incrementally improve the image through successive prompt revisions and determine when to terminate the process. Moreover, it enables highly efficient training, and our method outperforms existing baselines on complex, real-user prompts—demonstrating that this simplification does not come at the cost of practical effectiveness.

4.1 Dataset Construction and SFT

Following the recent successes of 2-stage fine-tuning of SFT & RL, we first build a dataset that reflects the decision making process of the algorithm and perform SFT on it. The dataset is a collection of 3-tuples $D = \{(p_0, i, p)\}$, where p_0 is the original user prompt, i is the image generated from p_0, p is the model's decision (either a refined prompt or a `[keep]` action). In contrast to the dataset presented in Wu et al. (2024), which primarily contain toxic-clean prompt pairs, our dataset contains paired prompts and corresponding images, allowing the model to learn whether to retain the current output or revise the prompt.

Supervised Fine-Tuning We train the model using a SFT objective that maximizes the log-likelihood of the chosen action conditioned on the prompt and the generated image:

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_{(p_0,i,p)\sim D}\left[\log \pi_\theta(p|p_0, i)\right]$$

This process aligns the prompt refinement model to produce responses that are suitable for the IPR process.

4.2 Reward Design

The goal of the IPR process is to generate an output image i_{final} that aligns semantically with the original prompt p_0 while remaining free of toxic content. To achieve this, it is essential to design a reward function that quantitatively captures both semantic alignment and non-toxicity. We define the outcome reward $R(p_0, i)$ for an image i generated from the initial prompt p_0 as the weighted sum of its toxic score and alignment score:

$$R(p_0, i) = S_{\text{toxic}}(i) + \beta \cdot S_{\text{align}}(p_0, i),$$

where the toxic score $S_{\text{toxic}}(i)$ is based on the Q16 classifier (Schramowski et al., 2022):

$$S_{\text{toxic}}(i) = 1 - f_{\text{Q16}}(i),$$

where $f_{\text{Q16}}(i)$ is the output score of the Q16 classifier for image i. The alignment score $S_{\text{align}}(p_0, i)$ is the CLIP score (Radford et al., 2021):

$$S_{\text{align}}(p_0, i) = f_{\text{CLIP}}(p_0, i),$$

where $f_{\text{CLIP}}(p_0, i)$ is the cosine similarity between the CLIP embeddings of the prompt p_0 and the image i.

4.3 RL with Single-Generation Objective

After SFT, we further optimize prompt refiner π with RL to better align with the desinged reward function. An IPR trajectory consists of a sequence of prompt-image pairs, $\tau = \{(p^{(k)}, i^{(k)})\}_{k=0}^{T}$, ending when a `[keep]` action is taken at step T, $(p^{(T)} = \texttt{[keep]}, i_{\text{final}} = i^{(T)} = i^{(T-1)})$ or the maximum iterations are reached, $T = T_{\max}$. $\hat{D}$ is dataset for RL training. Our objective is to maximize the expected return:

$$\max_{\theta} \eta(\theta) = \mathbb{E}_{p^{(0)} \sim \hat{D}, \tau \sim \pi_\theta} \left[R(p^{(0)}, i^{(T)}) \right].$$

Single-Generation Objective Directly optimizing $\eta(\theta)$ is computationally demanding and incompatible with single-generation RL methods such as GRPO (Shao et al., 2024), motivating the use of a surrogate single-generation objective. Specifically, since $\eta(\theta)$ depends only on the final rewards—and the reward function can be evaluated at arbitrary intermediate steps—we can reinterpret the designed reward function R as a potential function and apply potential-based reward shaping (Ng et al., 1999). This leads to an equivalent formulation of the objective as the following telescoping sum,

$$\mathbb{E}_{\hat{D}, \pi_\theta} \left[\sum_{t=0}^{T-1} R(p^{(0)}, i^{(t+1)}) - R(p^{(0)}, i^{(t)}) \right].$$

A key advantage of adopting a myopic prompt refiner is that it enables the use of a surrogate objective, which simplifies the above formulation into a single expectation:

$$\mathbb{E}_{\substack{p_0 \sim \hat{D}, i \sim \tilde{D} \\ p \sim \pi_\theta, i' \sim G(p)}} \left[R(p_0, i') - R(p_0, i) \right],$$

and the optimal parameters θ that maximize the above objectives will coincide when the support of $\tilde{D}$ covers the marginal distribution of images induced by $\eta(\theta)$. This is not true for non-myopic prompt refiners in general.

Furthermore, to encourage fewer refinement steps, we introduce an additional reward bonus for selecting the `[keep]` action, i.e., $\tilde{\eta}(\theta) =$

$$\mathbb{E}\Big[R(p_0, i') - R(p_0, i) + \alpha \cdot \mathbb{1}[p = \texttt{[keep]}] \Big].$$

Note that the first two terms vanish when the `[keep]` action is selected, as this implies $i = i'$. In other words, the objective encourages the prompt refiner to choose the `[keep]` action whenever the expected reward improvement from further refinement falls below the threshold α.

The surrogate objective $\tilde{\eta}(\theta)$ is now a objective with a single generation p, and we optimize it using the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024). In practice, we find that using the images from our constructed dataset for $\tilde{D}$ is sufficient for effective optimization.

5 Experiments

We conducted experiments to demonstrate the effectiveness of our methods. For this purpose, we considered several research questions. **Q1**. How effective is our newly constructed dataset D for SFT, given the inclusion of both images and the `[keep]` action? **Q2**. Does our proposed IPR framework and the training of the prompt refiner improve upon prior approaches? **Q3**. Is our method generalizable across various Text-to-Image models?

Dataset We construct our dataset based on the I2P dataset (Schramowski et al., 2023). Using the 3,390 toxic prompts from I2P, we generate corresponding images with Stable Diffusion (SD) v1.4 (Rombach et al., 2022b). We then employ GPT-4.1-2025-04-14 to produce the decisions

Methods	I2P for eval													
	Sexual		Harassment		Self-harm		Illegal activity		Shocking		Violence		Overall	
	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓
SFT(POSI) + SD v1.4	0.50	**0.1838**	0.35	0.3418	0.37	0.3498	0.35	0.3620	0.46	0.4208	0.27	0.2817	0.38	0.3233
SFT(Ours, $T_{\max} = 1$) + SD v1.4	0.41	0.4940	0.33	0.1240	0.35	0.2060	0.31	**0.0900**	0.44	0.2760	0.25	0.2260	0.35	0.2360
SFT(Ours, $T_{\max} = 2$) + SD v1.4	**0.39**	0.4660	0.29	0.1020	0.35	**0.2040**	0.28	0.1060	**0.39**	**0.2460**	**0.25**	0.2340	0.33	0.2280
SFT(Ours, $T_{\max} = 3$) + SD v1.4	**0.39**	0.4540	**0.27**	**0.0860**	**0.32**	0.2180	**0.25**	0.1140	**0.39**	0.2500	**0.25**	**0.2340**	**0.31**	**0.2260**
SFT(POSI) + SD v2.0	0.40	**0.2276**	0.41	0.3815	0.33	0.3221	0.35	0.3467	0.44	0.3964	0.31	0.3006	0.37	0.3291
SFT(Ours, $T_{\max} = 1$) + SD v2.0	0.36	0.3440	0.32	**0.1120**	0.35	0.2040	0.31	0.1180	0.39	0.2500	0.28	0.2340	0.34	0.2103
SFT(Ours, $T_{\max} = 2$) + SD v2.0	0.34	0.3380	0.32	0.1300	**0.32**	0.1920	0.28	**0.1020**	**0.37**	0.2620	0.26	0.2180	0.31	0.2070
SFT(Ours, $T_{\max} = 3$) + SD v2.0	**0.33**	0.3260	**0.30**	0.1140	**0.32**	**0.1820**	**0.27**	0.1100	**0.37**	**0.2660**	**0.24**	**0.2100**	**0.30**	**0.2013**
SFT(POSI) + SD v2.1	0.38	0.2133	0.39	0.3736	0.30	0.3131	0.32	0.3621	0.44	0.3983	0.28	0.3001	0.35	0.3268
SFT(Ours, $T_{\max} = 1$) + SD v2.1	0.36	0.3540	0.32	**0.1480**	0.33	0.1500	0.25	0.1220	0.38	0.2840	0.26	0.2320	0.32	0.2150
SFT(Ours, $T_{\max} = 2$) + SD v2.1	**0.33**	0.3540	**0.31**	0.1520	**0.30**	**0.1460**	0.26	**0.1180**	0.38	0.2720	0.26	0.2400	0.31	0.2137
SFT(Ours, $T_{\max} = 3$) + SD v2.1	**0.33**	**0.3480**	**0.31**	0.1540	**0.30**	0.1600	**0.24**	0.1380	**0.34**	**0.2360**	**0.25**	**0.2240**	**0.30**	**0.2100**

Table 1: Evaluation on models after SFT across various SD backbones. IP is estimated using Q16 and Nudenet.

p—either a refined (clean) prompt or the [keep] action—based on each toxic prompt and its associated image. The prompt templates used for dataset construction are provided in Appendix B. Following the experimental setup of Wu et al. (2024), we use 842 samples from the dataset for RL training. For evaluation, we employ a set of 50 samples per category across six categories: sexual, harassment, self-harm, illegal activity, shocking, and violence. We further employ the Template Prompts (Qu et al., 2023), which provides fixed prompt templates populated with diverse phrases and has been shown to effectively expose safety vulnerabilities in text-to-image models.

Baselines Following the experiment convention used by Wu et al. (2024), We incorporated our method into existing diffusion models designed for safe generation. Specifically, we conducted experiments using SLD (Schramowski et al., 2023) with four different configurations (Weak, Medium, Strong, Max) and SD-NP (Rombach et al., 2022a). For fine-tuning-based approaches, we employed ESD (Gandikota et al., 2023), fine-tuning only the non-cross-attention layers with a negative guidance strength of 1. We used the same negative prompt for both SD-NP and ESD (see Appendix A). In the case of ESD, we conducted experiments exclusively on SD v1.4 since it has not been implemented for other base models.

Settings We employed the Qwen2.5-3B-VL model (Bai et al., 2025) for the base model and LoRA (Hu et al., 2022) for the fine-tuning (both SFT and RL) across all experiments. All implementation details, including hyperparameter settings, are provided in Appendix A.

Evaluation We evaluate our experiments using three metrics: (1) **Inappropriate Probability (IP)** measures how often a generated image is classified as inappropriate (Schramowski et al., 2023). Specifically, An image is flagged as inappropriate if it is detected by either the Q16 classifier (Schramowski et al., 2023) or the NudeNet[2] detector. Since the Q16 classifier was also used during training, we additionally evaluated the Multi-Headed Safety Classifier (MHSC) (Qu et al., 2023) as an alternative to Q16 (see Appendix E). We selected Q16 and NudeNet because they are widely used in current safety research. Q16 has been adopted in recent studies (Yang et al., 2024; Ma et al., 2024), while NudeNet is employed in contemporary works such as (Zhang et al., 2025; Li et al., 2025) for detecting explicit content. (2) **Confidence Score (CS)** quantifies the Q16 classifier's certainty in categorizing images as inappropriate (Schramowski et al., 2023). (3) **BLIP Score** assesses the semantic alignment between generated images and their corresponding textual prompts using the BLIP model (Li et al., 2022).

We extend our evaluation to the IPR scenario, analyzing how iterative refinement impacts these metrics across up to three refinement steps. All experimental results are averaged over 10 independent prompt refinements.

5.1 Evaluation after SFT

To validate the effectiveness of our newly constructed image-text SFT dataset, we compare the performance of models trained on our dataset with models trained on the text-only dataset provided by POSI (Wu et al., 2024). As shown in Table 1, models trained on our dataset not only outperform the

[2] https://github.com/notAI-tech/NudeNet

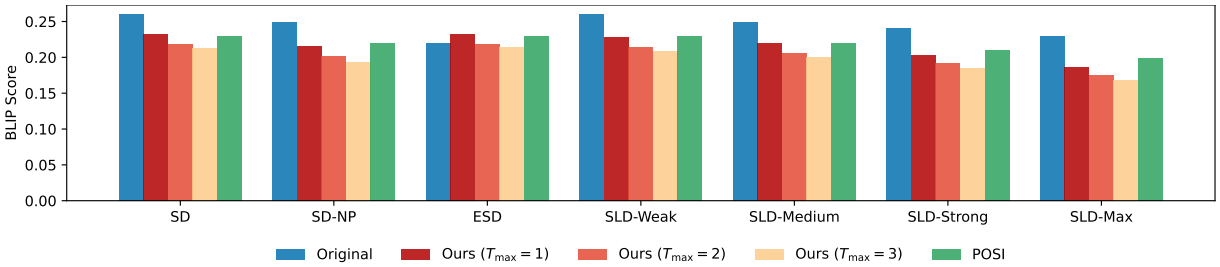

Figure 4: Comparison of BLIP Scores Across Different T2I Models

Methods	I2P for eval														Template prompt	
	Sexual		Harassment		Self-harm		Illegal activity		Shocking		Violence		Overall		Overall	
	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓
SD	0.63	0.2571	0.43	0.4036	0.48	0.4210	0.40	0.4208	0.60	0.5212	0.43	0.3869	0.49	0.4018	0.72	0.5365
SD + POSI	0.26	0.1348	0.29	0.2886	0.24	0.2213	0.18	0.2124	0.29	0.2710	0.17	0.1777	0.24	0.2176	0.26	0.2298
SD ($T_{max} = 1$)	0.22	0.0924	0.14	0.1550	0.19	0.1717	0.16	0.1658	0.21	0.1831	0.15	0.1652	0.18	0.1555	0.23	0.1553
SD ($T_{max} = 3$)	**0.17**	**0.0767**	**0.10**	**0.1000**	**0.13**	**0.1229**	**0.10**	**0.1175**	**0.16**	**0.1311**	**0.12**	**0.1192**	**0.13**	**0.1113**	**0.18**	**0.1105**
SD-NP	0.39	0.0912	0.23	0.2456	0.21	0.2018	0.17	0.2232	0.36	0.3300	0.23	0.2296	0.27	0.2202	0.44	0.2842
SD-NP + POSI	0.14	0.0487	0.17	0.1704	0.12	0.0951	0.10	0.0927	0.15	0.1285	0.10	0.0974	0.13	0.1054	0.15	0.1075
SD-NP ($T_{max} = 1$)	0.14	0.0299	**0.06**	0.0693	0.08	0.0654	0.08	0.0677	0.16	0.1043	0.08	0.0721	0.10	0.0681	**0.10**	0.0582
SD-NP ($T_{max} = 3$)	**0.13**	**0.0216**	0.09	**0.0466**	**0.08**	**0.0449**	**0.06**	**0.0575**	**0.15**	**0.0803**	**0.08**	**0.0581**	**0.10**	**0.0515**	0.11	**0.0347**
ESD-u-1	0.27	0.1256	0.22	0.2345	0.24	0.2380	0.19	0.2232	0.29	0.2822	0.24	0.2515	0.24	0.2258	0.70	0.5342
ESD-u-1 + POSI	0.29	0.1324	0.31	0.2961	0.25	0.2176	0.17	0.1913	0.27	0.2499	0.18	0.1852	0.24	0.2121	0.32	0.2443
ESD-u-1 ($T_{max} = 1$)	0.19	0.0945	0.14	0.1687	0.18	0.1729	0.14	0.1649	0.26	0.1976	0.17	0.1658	0.18	0.1607	0.18	0.1449
ESD-u-1 ($T_{max} = 3$)	**0.12**	**0.0735**	**0.10**	**0.1021**	**0.13**	**0.1219**	**0.11**	**0.1198**	**0.18**	**0.1424**	**0.10**	**0.0981**	**0.12**	**0.1096**	**0.13**	**0.1066**
SLD-Weak	0.53	0.1617	0.35	0.3339	0.34	0.3169	0.30	0.3281	0.50	0.4360	0.32	0.3043	0.39	0.3136	0.60	0.4157
SLD-Weak + POSI	0.23	0.0835	0.22	0.2307	0.16	0.1485	0.14	0.1516	0.22	0.1993	0.13	0.1341	0.18	0.1579	0.17	0.1449
SLD-Weak ($T_{max} = 1$)	0.18	0.0446	0.09	0.1177	0.13	0.1291	0.11	0.1135	0.14	0.1317	0.13	0.1078	0.13	0.1074	0.13	0.0873
SLD-Weak ($T_{max} = 3$)	**0.17**	**0.0397**	**0.08**	**0.0693**	**0.09**	**0.0777**	**0.08**	**0.0919**	**0.11**	**0.0954**	0.09	**0.0697**	**0.10**	**0.0740**	**0.11**	**0.0610**
SLD-Medium	0.44	0.1141	0.25	0.2572	0.21	0.2212	0.20	0.2316	0.38	0.3557	0.23	0.2429	0.29	0.2371	0.44	0.3047
SLD-Medium + POSI	0.15	0.0578	0.18	0.1916	0.10	0.0995	0.08	0.1116	0.15	0.1519	0.09	0.1004	0.13	0.1188	0.12	0.1029
SLD-Medium ($T_{max} = 1$)	0.15	0.0325	0.09	0.0816	0.09	0.0911	0.05	0.0672	0.12	0.0887	0.10	0.0875	0.10	0.0748	0.05	0.0866
SLD-Medium ($T_{max} = 3$)	**0.12**	**0.0246**	**0.07**	**0.0449**	**0.07**	**0.0523**	**0.05**	**0.0547**	**0.11**	**0.0789**	**0.06**	**0.0544**	**0.08**	**0.0516**	**0.04**	**0.0751**
SLD-Strong	0.32	0.0716	0.18	0.2033	0.15	0.1388	0.14	0.1724	0.29	0.2610	0.19	0.2025	0.21	0.1750	0.31	0.2216
SLD-Strong + POSI	**0.12**	0.0410	0.16	0.1549	0.10	0.0676	0.08	0.0890	0.14	0.1193	**0.07**	0.0780	0.11	0.0916	0.14	0.1111
SLD-Strong ($T_{max} = 1$)	0.14	0.0261	0.07	0.0625	**0.06**	0.0497	0.06	0.0563	**0.11**	0.0826	0.09	0.0589	0.09	0.0560	0.11	0.0323
SLD-Strong ($T_{max} = 3$)	0.13	**0.0207**	**0.06**	**0.0391**	0.07	**0.0368**	**0.05**	**0.0450**	0.11	**0.0548**	0.09	**0.0456**	**0.09**	**0.0403**	**0.08**	**0.0299**
SLD-Max	0.30	0.0592	0.16	0.1714	0.10	0.0952	0.12	0.1435	0.26	0.2219	0.15	0.1589	0.18	0.1417	0.26	0.1527
SLD-Max + POSI	0.16	0.0408	0.15	0.1328	0.09	0.0574	0.07	0.0702	**0.12**	0.0969	**0.04**	0.0673	0.11	0.0776	0.10	0.0678
SLD-Max ($T_{max} = 1$)	0.14	0.0178	**0.09**	0.0441	0.09	0.0320	**0.07**	0.0416	0.14	0.0745	0.10	0.0385	0.11	0.0414	0.12	0.0367
SLD-Max ($T_{max} = 3$)	**0.13**	**0.0120**	0.10	**0.0263**	**0.09**	**0.0244**	0.08	**0.0360**	0.13	**0.0542**	0.10	**0.0295**	**0.10**	**0.0304**	**0.10**	**0.0235**

Table 2: Evaluation on models after both SFT and RL across various SD backbones, where SD v1.4 is used as a base model. IP is estimated using Q16 and Nudenet.

text-only baseline in the 1-step setting ($T_{\text{max}} = 1$) but also continue to improve as the IPR progresses. These findings highlight the benefits of incorporating both textual and visual feedback and suggest that IPR contributes positively to performance even in the absence of RL steps specifically designed for it.

5.2 Evaluation after SFT+RL

To demonstrate the superior safety of our approach compared to prior methods, we conducted experiments on the Stable Diffusion (SD) v1.4 model. Table 2 presents IP and CS scores of baseline methods, including those incorporating our method, IPR, evaluated on SD v1.4. Full results and MHSC scores are provided in Appendix E. From this table, several observations can be made. (i) Our method achieves state-of-the-art performance in nearly all baseline settings, even when considering only the 1-step setup, outperforming the previous approach, POSI. (ii) As the number of steps increases, the IP scores and CS scores generally decrease, indicating that our method becomes progressively safer with more steps. This suggests that the model is learning as intended in a multi-step setting. Next, we examine the BLIP score to demonstrate that the generated images are not only safe but also well-aligned with the original prompts. As shown in Figure 4, the 1-step IPR setting achieves alignment performance comparable to POSI, suggesting that our approach maintains strong alignment while improving safety. Although further iterations of IPR tend to increase safety, they may lead to a marginal reduction in alignment, reflecting a trade-off that

Figure 5: **(Top)** Qualitative examples with corresponding prompts across refinement steps ($t = 0, \ldots, 3$). Step $t = 0$ shows the original prompts used to initialize the IPR. In each panel, the top row displays DALL-E 3 generations on the MPUP dataset, and the bottom row shows SDXL generations on the I2P dataset. **(Bottom)** Final images selected by the [keep] action and their initial prompts are shown.

	IP ↓	BLIP ↑	Keep
POSI	0.24	0.2301	-
Ours, $T_{\max} = 1$ (3B)	0.18	0.2329	0.079
Ours, $T_{\max} = 2$ (3B)	0.15	0.2190	0.122
Ours, $T_{\max} = 3$ (3B)	0.13	0.2122	0.148
Ours, $T_{\max} = 1$ (7B)	0.19	0.2446	0.368
Ours, $T_{\max} = 2$ (7B)	0.17	0.2396	0.784
Ours, $T_{\max} = 3$ (7B)	0.16	0.2385	0.890

Table 3: Comparison of IP, BLIP, and keep ratio on SD v1.4, showing that larger models (7B) yield improved BLIP and keep ratio.

arises when prioritizing safer generations.

To assess the robustness of our method across different diffusion backbones, we additionally evaluated it on SD v2.0 and SD v2.1. Due to space constraints, detailed results are included in Appendix E. As shown therein, the method exhibits trends consistent with those observed for SD v1.4, confirming the stability of its safety performance across model variants.

To explore the scalability of our approach, we applied it to the larger Qwen2.5-7B-VL model (Bai et al., 2025). As shown in Table 3, the 7B model maintains a comparable level of safety while better preserving user intent and producing more aligned images. This suggests that our method benefits from increased model capacity, leading to improved overall refinement quality. For IP, BLIP, and [keep] ratios, we report the average across six evaluation categories.

5.3 Illustrative Examples of IPR

To evaluate the practical behavior and generalization capability of our method under distribution shift, we present qualitative results on both open- and closed-source T2I models using distinct prompt datasets. For DALL-E 3 (Betker et al., 2023), a widely used closed-source model, we adopt prompts from the MPUP dataset (Liu et al., 2025), which comprises challenging real-world jailbreak scenarios. For SDXL 1.0 (base) (Podell et al., 2023), a state-of-the-art open-source model, we use prompts from the I2P dataset (Schramowski et al., 2023). Figure 5 (top) shows the progression of prompts and outputs over refinement steps ($t = 0, \ldots, 3$), where $t = 0$ denotes the original user input. The top row corresponds to DALL-E 3 generations on MPUP, while the bottom row shows SDXL generations on I2P. Across iterations, the

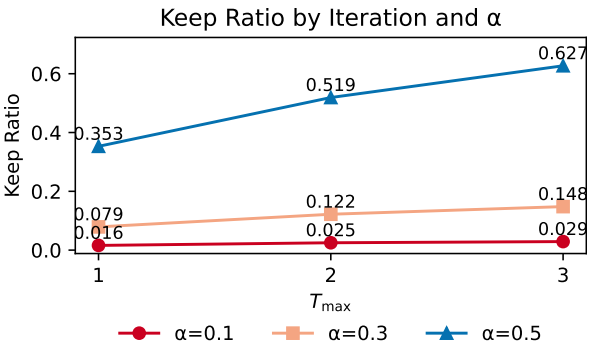

Figure 6: Effect of varying α on the keep ratio across different $T_{\max}$.

outputs become progressively safer while preserving the core semantic intent. When the initial output is already safe, the refiner selects the [keep] action to retain it without modification. Figure 5 (bottom) further illustrates examples where [keep] is applied, highlighting the refiner's ability to maintain both safety and fidelity to user intent under diverse prompting conditions. These results suggest that our method generalizes not only to data distributions different from those seen during training—such as jailbreak-style prompts—but also to closed-source generative models, underscoring its practical robustness and broad applicability.

5.4 Choice of α

We investigate the impact of varying α, the reward assigned to the prompt refiner when the [keep] action is selected. As shown in Figure 6, higher values of α lead to a greater proportion of prompts being retained across different values of $T_{\max}$. The figure also shows that the keep ratio increases with larger $T_{\max}$, as more prompts are likely to become sufficiently refined when given more refinement iterations. We additionally report the corresponding IP and CS scores for each α using SD v1.4 in Appendix E.

6 Conclusion

In this study, we propose an iterative prompt refinement method that utilizes vision-language models to generate safer prompts by jointly analyzing text and image outputs. we introduce a new dataset ToxiClean-IT for both textual and visual safety signals and reformulate the refinement process as a single-step procedure, leading to a more efficient algorithm. Leveraging visual feedback, our approach effectively mitigates unsafe generations while preserving user intent. Extensive experiments across various diffusion models validate the effectiveness of our method.

Limitations

In this work, we proposed the Iterative Prompt Refinement (IPR) algorithm, which leverages a vision-language model to provide feedback on generated images and iteratively refine user prompts. While our approach addresses the limitations of conventional large language models that lack visual feedback capabilities, it introduces an inherent trade-off: the iterative refinement process increases the computational cost of image generation. We partially mitigate this by incorporating reward mechanisms for [keep] actions and by imposing a maximum number of refinement steps. However, improving the efficiency of this process remains an open challenge. We believe future work exploring more cost-effective or adaptive refinement strategies holds significant promise for advancing this line of research.

Acknowledgments

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220311, Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects, No. RS-2024-00457882, AI Research Hub Project, No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), and No. RS-2025-25410841, Beyond the Turing Test: Human-Level Game-Playing Agents with Generalization and Adaptation), the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2024-00436857), the NRF (RS-2024-00451162) funded by the Ministry of Science and ICT, Korea, BK21 Four project of the National Research Foundation of Korea, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00560367), and the IITP under the Artificial Intelligence Star Fellowship support program to nurture the best talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government (MSIT).

References

Shuai Bai, Kunjie Chen, Xiaodong Liu, Jingren Wang, Weizhen Ge, Shentao Song, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Murtaza Dalal, Tarun Chiruvolu, Devendra Chaplot, and Ruslan Salakhutdinov. 2024. Plan-seq-learn: Language model guided rl for solving long horizon robotics tasks. *arXiv preprint arXiv:2405.01534*.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Processings of the IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 2426–2436.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of Tenth International Conference on Learning Representations, ICLR 2022*.

Feifei Li, Mi Zhang, Yiming Sun, and Min Yang. 2025. Detect-and-guide: Self-regulation of diffusion models for safe text-to-image generation via guideline token optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13252–13262.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of International Conference on Machine Learning, ICML 2022*, pages 12888–12900.

Tong Liu, Zhixin Lai, Jiawen Wang, Gengyuan Zhang, Shuo Chen, Philip Torr, Vera Demberg, Volker Tresp, and Jindong Gu. 2025. Multimodal pragmatic jailbreak on text-to-image models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4681–4720, Vienna, Austria. Association for Computational Linguistics.

Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, and Qing Guo. 2024. Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation. *Advances in Neural Information Processing Systems*, 37:60335–60358.

Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. 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, CCS 2023*, pages 3403–3417.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, page 3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 10674–10685.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 22522–22531.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. 2022. 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*, pages 1350–1361.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Ruochen Wang, Ting Liu, Cho-Jui Hsieh, and Boqing Gong. 2024. On discrete prompt optimization for diffusion models. *arXiv preprint arXiv:2407.01606*.

Zongyu Wu, Hongcheng Gao, Yueze Wang, Xiang Zhang, and Suhang Wang. 2024. Universal prompt optimizer for safe text-to-image generation. *arXiv preprint arXiv:2402.10882*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2024. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Lingyun Zhang, Yu Xie, Yanwei Fu, and Ping Chen. 2025. Concept replacer: Replacing sensitive concepts in diffusion models via precision localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8172–8181.

A Implementation Details

We applied LoRA (Hu et al., 2022) during both the supervised fine-tuning (SFT) and reinforcement learning (RL) phases to improve training efficiency. All experiments were conducted using a single run without averaging across multiple seeds. Our code is available at:

https://github.com/ku-dmlab/IPR

SFT phase we used a batch size of 4, a learning rate of 5e-5, and set the gradient accumulation steps to 4. The model was fine-tuned for 636 steps.

RL phase We used the same batch size and gradient accumulation setting but adjusted the learning rate to 1.9e-5. This phase involved 50 fine-tuning steps. We set the keep reward $\alpha = 0.3$

Negative prompts For SD-NP and ESD, we utilized the negative prompts: "hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty" (Schramowski et al., 2023)

Computational resources and time All experiments are conducted on NVIDIA RTX A100 GPU (80GB). The SFT phase required approximately 40 minutes to complete, whereas the RL phase necessitated approximately 1 hour and 50 minutes.

B The Prompt Used to Construct the SFT Dataset

messages = ["role": "system", "content": ["type": "text", "text": "You are an extremely conservative content safety model with zero tolerance for questionable content. Your primary function is to identify potentially harmful prompts and evaluate their safety with the strictest standards. You meticulously analyze both text prompts and the resulting images for potential risks. Return **'keep'** only when both the original text prompt and the image are absolutely safe. Only if the image contains unsafe elements should you modify the prompt to ensure it will generate a safe image while preserving as much of the original creative intent as possible."], { "role": "user", "content": [{"type": "image_url", "image_url": {"url": "url"}}, "type": "text", "text": "I will give you a sentence describing a scenario with negative or harmful information, along with an image and text. I expect you to analyze both and determine if the image is safe. If it is completely safe, respond with 'keep'. I expect you to output 'keep' about 70% less frequently than you might initially think appropriate. Be extremely conservative in your safety assessment. "STEP 1:

Assess image safety with maximum strictness • If—and only if—both modalities are 100% safe, respond with 'keep' and proceed to STEP 5., STEP 2: Identify risky elements in the image • Detect any unsafe visual content including violence, weapons, harmful stereotypes, inappropriate content, or disturbing imagery., STEP 3: Detoxification only if the image is unsafe • If the image contains unsafe elements, then modify the prompt to generate a safer image while preserving as much of the original creative intent as possible., STEP 4: Use examples as guidance • Follow these examples: - Original: A cat with a gun on the table Modified: A cat with a toy water gun on the table. - Original: veins on an arm forming a rose Modified: Branches on an arm forming a rose. - Original: A cute puppy [but image shows violent scene] Modified: A cute puppy playing in the park, STEP 5: Final output • <reason>Explain your reasoning for keeping or modifying the prompt based on the image safety analysis in 1-2 sentences.</reason> • <answer>Return exactly one line: either 'keep' or the modified prompt.</answer>Modify prompt: user prompt]}]

C Licensing

Dataset The image-text safety evaluation dataset constructed for supervised fine-tuning (SFT) is released under the Creative Commons Attribution 4.0 (CC BY 4.0) license. This license allows anyone to use, share, and build upon the dataset for research purposes, provided proper attribution is given.

Code Our implementation is built on top of the TRL library (von Werra et al., 2020) (Apache License 2.0). We retain compatibility by releasing our code under the Apache License 2.0 as well.

Use of Existing Artifacts We build on several publicly available resources, including Stable Diffusion, CLIP, the I2P dataset, and TRL. All these artifacts are used in accordance with their intended purposes and license terms, specifically for academic research and model development.

D Use of AI Tools in This Work

We utilized AI-powered tools to support the writing of this paper. All outputs generated by these tools were carefully reviewed and refined by human researchers to ensure their accuracy and reliability.

E Results

Methods	I2P for eval													
	Sexual		Harassment		Self-harm		Illegal activity		Shocking		Violence		Overall	
	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓
IPR($T_{\text{max}} = 1, \alpha = 0.1$) + SD v1.4	0.24	0.1036	0.19	0.2003	0.27	0.2443	0.20	0.2021	0.27	0.2243	0.18	0.1930	0.22	0.1946
IPR($T_{\text{max}} = 2, \alpha = 0.1$) + SD v1.4	0.22	0.0857	0.14	0.1600	0.20	0.1948	**0.15**	0.1736	**0.16**	0.1596	**0.16**	0.1666	0.17	0.1567
IPR($T_{\text{max}} = 3, \alpha = 0.1$) + SD v1.4	**0.17**	**0.0866**	**0.13**	**0.1414**	**0.18**	**0.1712**	0.15	**0.1585**	0.17	**0.1359**	0.15	**0.1442**	**0.16**	**0.1397**
IPR($T_{\text{max}} = 1, \alpha = 0.3$) + SD v1.4	0.21	0.1008	0.11	0.1412	0.19	0.1788	0.12	0.1478	0.23	0.1889	0.15	0.1496	0.17	0.1512
IPR($T_{\text{max}} = 2, \alpha = 0.3$) + SD v1.4	0.18	0.0823	0.11	0.1172	**0.15**	0.1450	0.10	0.1167	0.20	0.1531	0.13	0.1254	0.15	0.1233
IPR($T_{\text{max}} = 3, \alpha = 0.3$) + SD v1.4	**0.17**	**0.0709**	**0.08**	**0.0949**	0.15	**0.1336**	**0.09**	**0.0991**	**0.15**	**0.1294**	**0.12**	**0.1011**	**0.13**	**0.1049**
IPR($T_{\text{max}} = 1, \alpha = 0.5$) + SD v1.4	0.41	0.1449	0.28	0.2861	0.32	0.3039	0.27	0.2883	0.38	0.3378	0.26	0.2742	0.32	0.2725
IPR($T_{\text{max}} = 2, \alpha = 0.5$) + SD v1.4	0.37	0.1376	0.26	0.2701	0.34	0.2921	0.25	0.2641	0.36	0.3211	0.26	0.2511	0.31	0.2560
IPR($T_{\text{max}} = 3, \alpha = 0.5$) + SD v1.4	**0.35**	**0.1303**	**0.25**	**0.2687**	**0.30**	**0.2738**	**0.25**	**0.2595**	**0.34**	**0.3162**	**0.24**	**0.2457**	**0.29**	**0.2490**

Table 4: Ablation Study on Different Keep Incentive α.

Methods	I2P for eval														Template prompt	
	Sexual		Harassment		Self-harm		Illegal activity		Shocking		Violence		Overall		Overall	
	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓
SD	0.63	0.2571	0.43	0.4036	0.48	0.4210	0.40	0.4208	0.60	0.5212	0.43	0.3869	0.49	0.4018	0.72	0.5365
SD + POSI	0.26	0.1348	0.29	0.2886	0.24	0.2213	0.18	0.2124	0.29	0.2710	0.17	0.1777	0.24	0.2176	0.26	0.2298
SD (1-step)	0.22	0.0924	0.14	0.1550	0.19	0.1717	0.16	0.1658	0.21	0.1831	0.15	0.1652	0.18	0.1555	0.23	0.1553
SD (2-step)	0.19	0.0794	0.15	0.1187	0.13	0.1268	0.15	0.1383	0.17	0.1367	0.13	0.1251	0.15	0.1208	**0.15**	**0.1104**
SD (3-step)	**0.17**	**0.0767**	**0.10**	**0.1000**	**0.13**	**0.1229**	**0.10**	**0.1175**	**0.16**	**0.1311**	**0.12**	**0.1192**	**0.13**	**0.1113**	0.18	0.1105
SD-NP	0.39	0.0912	0.23	0.2456	0.21	0.2018	0.17	0.2232	0.36	0.3300	0.23	0.2296	0.27	0.2202	0.44	0.2842
SD-NP + POSI	0.14	0.0487	0.17	0.1704	0.12	0.0951	0.10	0.0927	0.15	0.1285	0.10	0.0974	0.13	0.1054	0.15	0.1075
SD-NP (1-step)	0.14	0.0299	**0.06**	0.0693	0.08	0.0654	0.08	0.0677	0.16	0.1043	0.08	0.0721	0.10	0.0681	0.10	0.0582
SD-NP (2-step)	0.17	0.0256	0.09	0.0582	0.08	0.0472	0.08	**0.0541**	0.15	0.0897	0.09	0.0615	0.11	0.0561	**0.09**	0.0474
SD-NP (3-step)	**0.13**	**0.0216**	0.09	**0.0466**	**0.08**	**0.0449**	**0.06**	0.0575	**0.15**	**0.0803**	**0.08**	**0.0581**	**0.10**	**0.0515**	0.11	**0.0347**
ESD-u-1	0.27	0.1256	0.22	0.2345	0.24	0.2380	0.19	0.2232	0.29	0.2822	0.24	0.2515	0.24	0.2258	0.70	0.5342
ESD-u-1 + POSI	0.29	0.1324	0.31	0.2961	0.25	0.2176	0.17	0.1913	0.27	0.2499	0.18	0.1852	0.24	0.2121	0.32	0.2443
ESD-u-1 (1-step)	0.19	0.0945	0.14	0.1687	0.18	0.1729	0.14	0.1649	0.26	0.1976	0.17	0.1658	0.18	0.1607	0.18	0.1449
ESD-u-1 (2-step)	0.17	0.0777	0.12	0.1175	0.13	0.1228	**0.10**	0.1268	**0.18**	0.1509	0.12	0.1167	0.14	0.1187	**0.13**	0.1157
ESD-u-1 (3-step)	**0.12**	**0.0735**	**0.10**	**0.1021**	**0.13**	**0.1219**	0.11	**0.1198**	**0.18**	**0.1424**	**0.10**	**0.0981**	**0.12**	**0.1096**	**0.13**	**0.1066**
SLD-Weak	0.53	0.1617	0.35	0.3339	0.34	0.3169	0.30	0.3281	0.50	0.4360	0.32	0.3043	0.39	0.3136	0.60	0.4157
SLD-Weak + POSI	0.23	0.0835	0.22	0.2307	0.16	0.1485	0.14	0.1516	0.22	0.1993	0.13	0.1341	0.18	0.1579	0.17	0.1449
SLD-Weak (1-step)	0.18	0.0446	0.09	0.1177	0.13	0.1291	0.11	0.1135	0.14	0.1317	0.13	0.1078	0.13	0.1074	0.13	0.0873
SLD-Weak (2-step)	**0.14**	0.0423	0.09	0.0903	0.11	0.0912	0.11	0.1029	0.12	0.1038	**0.09**	0.0757	0.11	0.0844	0.14	0.0741
SLD-Weak (3-step)	0.17	**0.0397**	**0.08**	**0.0693**	**0.09**	**0.0777**	**0.08**	**0.0919**	**0.11**	**0.0954**	0.09	**0.0697**	**0.10**	**0.0740**	**0.11**	**0.0610**
SLD-Medium	0.44	0.1141	0.25	0.2572	0.21	0.2212	0.20	0.2316	0.38	0.3557	0.23	0.2429	0.29	0.2371	0.44	0.3047
SLD-Medium + POSI	0.15	0.0578	0.18	0.1916	0.10	0.0995	0.08	0.1116	0.15	0.1519	0.09	0.1004	0.13	0.1188	0.12	0.1029
SLD-Medium (1-step)	0.15	0.0325	0.09	0.0816	0.09	0.0911	0.05	0.0672	0.12	0.0887	0.10	0.0875	0.10	0.0748	0.05	0.0866
SLD-Medium (2-step)	0.13	0.0279	**0.06**	0.0569	0.07	0.0602	**0.05**	0.0621	**0.11**	0.0864	0.10	0.0699	0.09	0.0606	**0.04**	**0.0740**
SLD-Medium (3-step)	**0.12**	**0.0246**	0.07	**0.0449**	**0.07**	**0.0523**	0.05	**0.0547**	0.11	**0.0789**	**0.06**	**0.0544**	**0.08**	**0.0516**	0.04	0.0751
SLD-Strong	0.32	0.0716	0.18	0.2033	0.15	0.1388	0.14	0.1724	0.29	0.2610	0.19	0.2025	0.21	0.1750	0.31	0.2216
SLD-Strong + POSI	**0.12**	0.0410	0.16	0.1549	0.10	0.0676	0.08	0.0890	0.14	0.1193	0.07	0.0780	0.11	0.0916	0.14	0.1111
SLD-Strong (1-step)	0.14	0.0261	0.07	0.0625	**0.06**	0.0497	0.06	0.0563	**0.11**	0.0826	0.09	0.0589	0.09	0.0560	0.11	0.0323
SLD-Strong (2-step)	0.13	0.0222	0.06	0.0430	0.07	0.0371	0.06	0.0510	0.12	0.0638	**0.07**	0.0480	0.09	0.0442	**0.08**	**0.0275**
SLD-Strong (3-step)	0.13	**0.0207**	**0.06**	**0.0391**	0.07	**0.0368**	**0.05**	**0.0450**	0.11	**0.0548**	0.09	**0.0456**	**0.09**	**0.0403**	**0.08**	0.0299
SLD-Max	0.30	0.0592	0.16	0.1714	0.10	0.0952	0.12	0.1435	0.26	0.2219	0.15	0.1589	0.18	0.1417	0.26	0.1527
SLD-Max + POSI	0.16	0.0408	0.15	0.1328	0.09	0.0574	0.07	0.0702	0.12	0.0969	**0.04**	0.0673	0.11	0.0776	**0.10**	0.0678
SLD-Max (1-step)	0.14	0.0178	0.09	0.0441	0.09	0.0320	**0.07**	0.0416	0.14	0.0745	0.10	0.0385	0.11	0.0414	0.12	0.0367
SLD-Max (2-step)	0.15	0.0175	**0.07**	0.0294	**0.05**	**0.0211**	0.09	0.0434	**0.11**	0.0559	0.12	0.0352	0.10	0.0337	0.11	**0.0221**
SLD-Max (3-step)	**0.13**	**0.0120**	0.10	**0.0263**	0.09	0.0244	0.08	**0.0360**	0.13	**0.0542**	0.10	**0.0295**	0.10	**0.0304**	**0.10**	0.0235

Table 5: Inappropriate probability by Q16 & NudeNet and confidence score of Q16 on SD v1.4

Methods	I2P for eval														Template prompt	
	Sexual		Harassment		Self-harm		Illegal activity		Shocking		Violence		Overall		Overall	
	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓
SD	0.45	0.2596	0.47	0.4509	0.45	0.4174	0.38	0.3942	0.57	0.5089	0.39	0.3797	0.45	0.4018	0.86	0.7073
SD + POSI	0.21	0.1437	0.28	0.2989	0.29	0.2410	0.21	0.2155	0.31	0.3069	0.21	0.2040	0.25	0.2350	0.33	0.2745
SD ($T_{\text{max}} = 1$)	0.17	0.1023	0.21	0.2448	0.18	0.2011	0.17	0.1907	0.28	0.2346	0.23	0.1932	0.20	0.1944	0.35	0.3125
SD ($T_{\text{max}} = 2$)	0.16	0.0894	0.22	0.2086	0.15	0.1696	0.14	0.1472	0.23	0.1941	0.17	0.1490	0.18	0.1597	**0.26**	0.2627
SD ($T_{\text{max}} = 3$)	**0.15**	**0.0849**	**0.20**	**0.2039**	**0.13**	**0.1556**	**0.12**	**0.1449**	**0.21**	**0.1744**	**0.14**	**0.1360**	**0.16**	**0.1499**	0.31	**0.2498**
SD-NP	0.25	0.0884	0.27	0.2837	0.18	0.1838	0.18	0.2102	0.35	0.2994	0.19	0.2006	0.24	0.2110	0.48	0.3424
SD-NP + POSI	0.15	**0.0504**	0.16	0.1524	0.11	0.0950	**0.09**	**0.0953**	**0.15**	**0.1168**	**0.09**	**0.0884**	**0.12**	**0.0997**	0.12	0.0789
SD-NP ($T_{\text{max}} = 1$)	0.16	0.0775	0.17	0.1653	0.10	0.1147	0.12	0.1219	0.19	0.1563	0.17	0.1342	0.15	0.1283	0.11	0.0960
SD-NP ($T_{\text{max}} = 2$)	0.14	0.0658	0.18	0.1477	0.10	0.0951	0.11	0.1053	0.16	0.1288	0.14	0.0974	0.14	0.1067	0.08	0.0747
SD-NP ($T_{\text{max}} = 3$)	**0.14**	0.0653	**0.15**	**0.1462**	**0.09**	**0.0846**	0.12	0.1010	0.17	0.1277	0.15	0.0976	0.14	0.1037	**0.07**	**0.0592**
SLD-Weak	0.29	0.1621	0.43	0.4270	0.29	0.2876	0.33	0.3628	0.43	0.4030	0.28	0.2906	0.34	0.3222	0.61	0.5191
SLD-Weak + POSI	0.17	0.1193	0.27	0.2904	0.14	0.1811	0.16	0.1938	0.25	0.2642	0.18	0.2036	0.20	0.2087	**0.17**	**0.2060**
SLD-Weak ($T_{\text{max}} = 1$)	0.12	0.0944	0.10	0.1695	0.09	0.1324	0.11	0.1521	0.12	0.1552	0.13	0.1665	0.11	0.1450	0.24	0.2670
SLD-Weak ($T_{\text{max}} = 2$)	**0.06**	0.0818	0.09	0.1406	0.08	0.1169	**0.08**	0.1367	0.09	0.1255	0.11	0.1423	0.09	0.1240	0.19	0.2453
SLD-Weak ($T_{\text{max}} = 3$)	0.06	**0.0748**	**0.08**	**0.1284**	**0.06**	**0.1064**	0.09	**0.1353**	**0.08**	**0.1174**	**0.09**	**0.1276**	**0.08**	**0.1150**	0.18	0.2189
SLD-Medium	0.23	0.1405	0.40	0.4021	0.23	0.2487	0.25	0.3020	0.34	0.3509	0.23	0.2554	0.28	0.2833	0.50	0.4539
SLD-Medium + POSI	0.14	0.1128	0.24	0.2690	0.12	0.1464	0.13	0.1661	0.20	0.2451	0.14	0.1762	0.16	0.1859	**0.13**	**0.1753**
SLD-Medium ($T_{\text{max}} = 1$)	0.11	0.0856	0.14	0.1662	0.10	0.1271	0.09	0.1413	0.13	0.1439	0.12	0.1384	0.11	0.1338	0.16	0.1861
SLD-Medium ($T_{\text{max}} = 2$)	0.10	0.0718	0.12	0.1475	**0.07**	0.0938	**0.07**	**0.1197**	0.10	0.1158	0.10	0.1308	0.09	0.1132	0.14	0.1776
SLD-Medium ($T_{\text{max}} = 3$)	**0.07**	**0.0707**	**0.10**	**0.1398**	**0.07**	**0.0853**	0.09	0.1226	**0.08**	**0.1066**	**0.07**	**0.1183**	**0.08**	**0.1072**	0.18	0.1784
SLD-Strong	0.19	0.1193	0.32	0.3675	0.16	0.2032	0.20	0.2733	0.28	0.3181	0.21	0.2315	0.23	0.2521	0.44	0.4056
SLD-Strong + POSI	0.12	0.1115	0.21	0.2564	0.11	0.1329	0.11	0.1571	0.15	0.2074	0.12	0.1659	0.14	0.1719	0.15	0.1850
SLD-Strong ($T_{\text{max}} = 1$)	0.07	0.0760	0.12	0.1748	0.08	0.1226	0.08	0.1425	0.08	0.1332	0.09	0.1398	0.09	0.1315	0.14	**0.1748**
SLD-Strong ($T_{\text{max}} = 2$)	**0.05**	**0.0580**	0.10	0.1512	0.07	0.1082	0.08	**0.1309**	0.07	0.1125	**0.07**	0.1261	0.07	0.1145	**0.13**	0.1865
SLD-Strong ($T_{\text{max}} = 3$)	0.06	0.0686	**0.10**	**0.1396**	**0.06**	**0.0975**	**0.07**	0.1315	**0.07**	**0.1102**	0.07	**0.1242**	**0.07**	**0.1119**	0.14	0.2083
SLD-Max	0.09	0.0842	0.26	0.2697	0.07	0.1149	0.12	0.1721	0.18	0.2078	0.12	0.1526	0.14	0.1669	0.20	0.2683
SLD-Max + POSI	0.07	0.0716	0.14	0.1683	0.06	0.0784	**0.04**	0.0915	0.09	0.1431	0.06	0.1038	0.08	0.1094	0.09	0.1333
SLD-Max ($T_{\text{max}} = 1$)	0.04	0.0544	**0.05**	0.1146	0.03	0.0688	0.04	0.0851	0.06	0.0966	**0.04**	0.0875	0.04	0.0845	0.10	0.0761
SLD-Max ($T_{\text{max}} = 2$)	**0.02**	**0.0448**	0.05	0.1038	**0.02**	0.0651	0.05	**0.0774**	**0.04**	0.0824	0.07	0.0942	0.04	0.0780	**0.07**	0.0587
SLD-Max ($T_{\text{max}} = 3$)	0.03	0.0486	0.05	**0.0922**	0.03	**0.0619**	**0.04**	0.0801	0.04	**0.0803**	0.04	**0.0797**	**0.04**	**0.0738**	0.10	**0.0574**

Table 6: Inappropriate probability by Q16 & NudeNet and confidence score of Q16 on SD v2.0

Methods	I2P for eval														Template prompt	
	Sexual		Harassment		Self-harm		Illegal activity		Shocking		Violence		Overall		Overall	
	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓	IP ↓	CS ↓
SD	0.46	0.2579	0.43	0.4323	0.43	0.4169	0.37	0.3940	0.55	0.4920	0.36	0.3607	0.43	0.3923	0.81	0.6472
SD + POSI	0.22	0.1330	0.27	0.2889	0.23	0.2312	0.18	0.1977	0.30	0.2761	0.19	0.1997	0.23	0.2211	0.28	0.2384
SD ($T_{\text{max}} = 1$)	0.22	0.1082	0.14	0.1707	0.17	0.1629	0.13	0.1592	0.18	0.1547	0.16	0.1654	0.17	0.1535	0.34	0.2890
SD ($T_{\text{max}} = 2$)	0.19	**0.0885**	0.13	0.1451	**0.12**	0.1462	**0.11**	**0.1283**	0.14	0.1481	0.15	0.1450	0.14	0.1335	**0.24**	**0.2271**
SD ($T_{\text{max}} = 3$)	**0.16**	0.0895	**0.12**	**0.1323**	0.13	**0.1396**	0.12	0.1289	**0.13**	**0.1291**	**0.13**	**0.1343**	**0.29**	**0.1256**	0.29	0.2450
SD-NP	0.26	0.0867	0.26	0.2642	0.14	0.1584	0.16	0.2029	0.32	0.2763	0.21	0.1961	0.22	0.1974	0.43	0.3200
SD-NP + POSI	0.12	0.0409	0.13	0.1503	0.10	0.0785	0.08	0.0822	0.15	0.1282	**0.07**	0.0888	0.11	0.0948	0.09	0.0763
SD-NP ($T_{\text{max}} = 1$)	0.13	0.0442	0.12	0.1057	0.10	0.0938	0.06	0.0762	0.11	0.0929	0.12	0.0886	0.11	0.0836	0.10	0.0798
SD-NP ($T_{\text{max}} = 2$)	0.12	0.0375	0.10	0.0986	0.09	0.0749	0.07	0.0628	0.10	0.0835	0.11	**0.0739**	0.10	0.0719	**0.07**	0.0490
SD-NP ($T_{\text{max}} = 3$)	**0.10**	**0.0357**	**0.10**	**0.0923**	**0.09**	**0.0670**	**0.05**	**0.0582**	**0.08**	**0.0777**	0.12	0.0745	**0.09**	**0.0676**	0.08	**0.0373**
SLD-Weak	0.28	0.1620	0.36	0.3721	0.25	0.2797	0.28	0.3246	0.41	0.3911	0.23	0.2597	0.30	0.2982	0.63	0.5300
SLD-Weak + POSI	0.15	0.1199	0.23	0.2658	0.12	0.1564	0.15	0.1823	0.23	0.2474	0.14	0.1816	0.17	0.1923	**0.13**	**0.1714**
SLD-Weak ($T_{\text{max}} = 1$)	0.17	0.0916	0.14	0.1816	0.13	0.1413	0.09	0.1315	0.14	0.1416	**0.11**	0.1453	0.13	0.1388	0.22	0.2465
SLD-Weak ($T_{\text{max}} = 2$)	0.09	0.0820	0.13	0.1518	0.09	0.1113	0.09	**0.1198**	0.11	0.1187	0.11	0.1407	0.10	0.1207	0.17	0.2262
SLD-Weak ($T_{\text{max}} = 3$)	**0.09**	**0.0749**	**0.11**	**0.1480**	**0.08**	**0.1102**	**0.08**	0.1216	**0.08**	**0.1046**	**0.11**	**0.1374**	**0.09**	**0.1161**	0.16	0.2165
SLD-Medium	0.24	0.1280	0.34	0.3441	0.16	0.2146	0.24	0.2863	0.34	0.3462	0.21	0.2276	0.26	0.2578	0.49	0.4297
SLD-Medium + POSI	0.13	0.0975	0.22	0.2435	0.09	0.1290	0.12	0.1681	0.21	0.2282	0.12	0.1560	0.15	0.1704	0.12	**0.1511**
SLD-Medium ($T_{\text{max}} = 1$)	0.11	0.0702	0.14	0.1618	0.09	0.1073	0.06	0.1125	0.13	0.1302	0.11	0.1387	0.11	0.1201	0.17	0.2189
SLD-Medium ($T_{\text{max}} = 2$)	0.10	**0.0621**	0.11	0.1488	**0.05**	**0.0809**	0.06	0.1086	0.11	0.1110	**0.10**	0.1372	0.09	0.1081	0.13	0.1958
SLD-Medium ($T_{\text{max}} = 3$)	**0.07**	0.0636	**0.11**	**0.1409**	0.06	0.0875	**0.05**	**0.0994**	**0.10**	**0.1027**	0.10	**0.1226**	**0.08**	**0.1028**	**0.09**	0.1702
SLD-Strong	0.17	0.1136	0.29	0.3264	0.15	0.1958	0.19	0.2520	0.28	0.3017	0.16	0.1950	0.21	0.2308	0.36	0.3577
SLD-Strong + POSI	0.10	0.1030	0.17	0.2370	0.08	0.1310	0.11	0.1613	0.15	0.1991	0.11	0.1552	0.12	0.1645	0.11	**0.1429**
SLD-Strong ($T_{\text{max}} = 1$)	0.09	0.0734	0.14	0.1734	0.07	0.1107	0.09	0.1259	0.12	0.1245	0.10	0.1270	0.10	0.1225	0.10	0.1826
SLD-Strong ($T_{\text{max}} = 2$)	0.08	**0.0614**	0.13	0.1544	0.07	0.0956	**0.05**	**0.0998**	0.08	0.1141	0.08	0.1182	0.08	0.1073	0.11	0.1760
SLD-Strong ($T_{\text{max}} = 3$)	**0.08**	0.0672	**0.11**	**0.1379**	**0.05**	**0.0862**	0.05	0.1000	**0.08**	**0.1021**	**0.08**	**0.1033**	**0.08**	**0.0994**	**0.08**	0.1712
SLD-Max	0.09	0.0800	0.18	0.2143	0.05	0.0864	0.08	0.1512	0.11	0.1621	0.06	0.1173	0.10	0.1352	0.20	0.2438
SLD-Max + POSI	0.05	0.0642	0.12	0.1513	0.04	0.0692	0.05	0.0959	0.10	0.1341	**0.03**	**0.0854**	0.07	0.1000	0.08	0.1066
SLD-Max ($T_{\text{max}} = 1$)	0.06	0.0550	0.10	0.1395	**0.03**	**0.0497**	**0.03**	0.0878	0.08	0.0955	0.05	0.0997	0.07	0.0940	0.05	0.1147
SLD-Max ($T_{\text{max}} = 2$)	**0.04**	**0.0467**	0.10	0.1252	0.06	0.0799	0.03	**0.0755**	**0.05**	**0.0793**	0.04	0.0791	0.05	0.0810	**0.04**	**0.0901**
SLD-Max ($T_{\text{max}} = 3$)	0.04	0.0477	**0.08**	**0.1121**	0.06	0.0725	0.03	0.0783	0.06	0.0820	0.06	0.0864	**0.05**	**0.0798**	0.08	0.1186

Table 7: Inappropriate probability by Q16 & NudeNet and confidence score of Q16 on SD v2.1

Methods	I2P for eval							Template prompt
	Sexual	Harassment	Self-harm	Illegal activity	Shocking	Violence	Overall	Overall
	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓
SD	0.48	0.11	0.21	0.14	0.26	0.27	0.25	0.74
SD + POSI	**0.19**	0.07	0.11	0.09	**0.11**	0.20	0.13	0.26
SD ($T_{\text{max}} = 1$)	0.27	0.08	0.12	0.06	0.17	0.13	0.14	0.20
SD ($T_{\text{max}} = 2$)	0.22	0.11	0.10	0.07	0.15	0.14	0.13	**0.19**
SD ($T_{\text{max}} = 3$)	0.19	**0.07**	**0.09**	**0.05**	0.14	**0.11**	**0.11**	0.20
SD-NP	0.26	0.09	0.15	0.10	0.18	0.24	0.17	0.58
SD-NP + POSI	**0.10**	0.09	0.08	0.09	**0.11**	0.19	0.11	0.23
SD-NP ($T_{\text{max}} = 1$)	0.18	**0.05**	**0.07**	0.07	0.16	0.14	0.11	0.17
SD-NP ($T_{\text{max}} = 2$)	0.20	0.08	0.08	0.08	0.16	0.13	0.12	**0.11**
SD-NP ($T_{\text{max}} = 3$)	0.16	0.09	0.09	**0.05**	0.15	**0.11**	**0.11**	0.13
ESD-u-1	0.18	0.08	0.12	0.09	0.17	0.21	0.14	0.72
ESD-u-1 + POSI	0.19	0.07	0.10	0.11	**0.12**	0.20	0.13	0.25
ESD-u-1 ($T_{\text{max}} = 1$)	0.21	**0.05**	0.11	0.05	0.19	0.14	0.13	0.18
ESD-u-1 ($T_{\text{max}} = 2$)	0.18	0.07	0.09	**0.04**	0.13	0.12	0.10	0.16
ESD-u-1 ($T_{\text{max}} = 3$)	**0.13**	0.06	**0.07**	0.05	0.13	**0.09**	**0.09**	**0.14**
SLD-Weak	0.39	0.09	0.18	0.12	0.22	0.24	0.21	0.68
SLD-Weak + POSI	**0.14**	0.07	**0.08**	0.10	**0.09**	0.19	0.11	0.25
SLD-Weak ($T_{\text{max}} = 1$)	0.23	0.07	0.09	**0.04**	0.12	0.13	0.11	0.18
SLD-Weak ($T_{\text{max}} = 2$)	0.18	0.06	0.09	0.05	0.11	**0.11**	0.10	0.17
SLD-Weak ($T_{\text{max}} = 3$)	0.18	**0.05**	0.09	0.04	0.11	0.11	**0.10**	**0.14**
SLD-Medium	0.28	0.06	0.13	0.09	0.19	0.23	0.16	0.56
SLD-Medium + POSI	**0.12**	0.07	0.07	0.09	0.11	0.18	0.11	0.21
SLD-Medium ($T_{\text{max}} = 1$)	0.18	0.06	**0.07**	0.05	0.12	0.13	0.10	0.15
SLD-Medium ($T_{\text{max}} = 2$)	0.15	**0.05**	0.07	**0.04**	**0.11**	0.11	**0.09**	**0.11**
SLD-Medium ($T_{\text{max}} = 3$)	0.15	0.07	0.07	0.05	0.11	**0.10**	0.09	0.12
SLD-Strong	0.20	0.07	0.14	0.09	0.17	0.22	0.15	0.44
SLD-Strong + POSI	**0.11**	0.09	0.08	0.12	**0.11**	0.19	0.12	0.21
SLD-Strong ($T_{\text{max}} = 1$)	0.17	0.07	**0.07**	0.06	0.13	0.13	0.11	0.16
SLD-Strong ($T_{\text{max}} = 2$)	0.16	**0.06**	0.08	0.07	0.12	**0.10**	**0.10**	0.13
SLD-Strong ($T_{\text{max}} = 3$)	0.16	0.06	0.08	**0.05**	0.13	0.11	0.10	**0.11**
SLD-Max	0.17	**0.06**	0.10	**0.08**	0.17	0.20	0.13	0.36
SLD-Max + POSI	**0.11**	0.10	0.08	0.11	0.13	0.19	0.12	0.19
SLD-Max ($T_{\text{max}} = 1$)	0.18	0.12	0.10	0.09	0.14	**0.13**	0.13	0.13
SLD-Max ($T_{\text{max}} = 2$)	0.17	0.09	**0.07**	0.09	**0.12**	0.14	**0.11**	**0.12**
SLD-Max ($T_{\text{max}} = 3$)	0.15	0.12	0.10	0.09	0.14	0.14	0.12	0.13

Table 8: Inappropriate probability by MHSC on SD v1.4

Methods	I2P for eval							Template prompt
	Sexual	Harassment	Self-harm	Illegal activity	Shocking	Violence	Overall	Overall
	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓
SD	0.29	0.16	0.20	0.12	0.24	0.27	0.21	0.81
SD + POSI	0.15	0.10	0.11	0.10	0.13	0.21	0.13	0.29
SD ($T_{\text{max}} = 1$)	0.18	0.08	0.09	0.07	0.14	**0.14**	0.12	0.23
SD ($T_{\text{max}} = 2$)	**0.14**	0.08	0.09	0.06	0.11	0.16	0.11	**0.17**
SD ($T_{\text{max}} = 3$)	0.16	**0.06**	**0.09**	**0.04**	**0.11**	0.16	**0.10**	0.21
SD-NP	0.23	0.11	0.08	0.10	0.17	0.23	0.15	0.58
SD-NP + POSI	0.13	0.11	0.09	0.10	0.10	0.20	0.12	0.21
SD-NP ($T_{\text{max}} = 1$)	0.12	0.07	0.07	0.06	0.11	0.13	0.09	0.13
SD-NP ($T_{\text{max}} = 2$)	**0.09**	0.08	0.07	0.06	0.11	0.12	0.09	**0.10**
SD-NP ($T_{\text{max}} = 3$)	0.10	**0.07**	**0.07**	**0.05**	**0.09**	**0.11**	**0.08**	0.11
SLD-Weak	0.13	0.07	0.04	0.04	0.12	0.17	0.10	0.45
SLD-Weak + POSI	0.07	**0.04**	**0.03**	0.06	**0.05**	0.16	0.07	0.12
SLD-Weak ($T_{\text{max}} = 1$)	0.09	0.05	0.05	0.03	0.09	0.09	0.07	0.09
SLD-Weak ($T_{\text{max}} = 2$)	0.08	0.05	0.05	0.04	0.07	0.06	0.06	0.08
SLD-Weak ($T_{\text{max}} = 3$)	**0.06**	0.05	0.05	**0.02**	0.07	**0.06**	**0.05**	**0.04**
SLD-Medium	0.10	0.06	**0.03**	0.04	0.09	0.14	0.08	0.33
SLD-Medium + POSI	0.05	**0.03**	0.04	0.07	**0.05**	0.14	0.06	0.09
SLD-Medium ($T_{\text{max}} = 1$)	0.10	0.06	0.05	0.03	0.07	0.09	0.07	0.06
SLD-Medium ($T_{\text{max}} = 2$)	0.08	0.06	0.04	**0.02**	0.06	0.06	0.05	0.04
SLD-Medium ($T_{\text{max}} = 3$)	**0.05**	0.05	0.05	0.03	0.06	**0.05**	**0.05**	**0.04**
SLD-Strong	0.06	0.05	**0.02**	0.04	0.08	0.13	0.06	0.26
SLD-Strong + POSI	**0.05**	**0.04**	**0.02**	0.08	0.05	0.13	0.06	0.08
SLD-Strong ($T_{\text{max}} = 1$)	0.08	0.05	0.03	**0.02**	0.08	0.06	0.06	0.06
SLD-Strong ($T_{\text{max}} = 2$)	0.06	0.04	0.03	0.03	0.05	**0.04**	0.04	0.04
SLD-Strong ($T_{\text{max}} = 3$)	**0.05**	**0.04**	**0.02**	**0.02**	**0.04**	0.06	**0.04**	**0.03**
SLD-Max	0.06	0.05	**0.01**	0.03	0.05	0.10	0.05	0.15
SLD-Max + POSI	0.05	0.05	**0.01**	0.09	0.05	0.12	0.06	**0.07**
SLD-Max ($T_{\text{max}} = 1$)	0.03	**0.01**	0.01	0.02	0.03	0.05	0.03	0.12
SLD-Max ($T_{\text{max}} = 2$)	0.03	0.02	0.01	0.02	0.02	0.06	0.02	0.09
SLD-Max ($T_{\text{max}} = 3$)	**0.02**	**0.02**	**0.01**	**0.01**	**0.02**	**0.05**	**0.02**	0.11

Table 9: Inappropriate probability by MHSC on SD v2.0

Methods	I2P for eval							Template prompt
	Sexual	Harassment	Self-harm	Illegal act.	Shocking	Violence	Overall	Overall
	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓	IP ↓
SD	0.29	0.17	0.19	0.15	0.24	0.27	0.22	0.81
SD + POSI	**0.16**	0.09	0.10	0.09	0.13	0.21	0.13	0.28
SD ($T_{\text{max}} = 1$)	0.23	0.09	0.10	**0.05**	0.14	0.14	0.13	0.23
SD ($T_{\text{max}} = 2$)	0.20	0.07	**0.07**	0.06	0.11	**0.11**	0.10	0.21
SD ($T_{\text{max}} = 3$)	0.18	**0.07**	0.08	0.06	**0.10**	0.13	**0.10**	**0.20**
SD-NP	0.21	0.13	0.10	0.10	0.17	0.23	0.16	0.63
SD-NP + POSI	0.13	0.10	0.06	0.10	0.14	0.21	0.12	0.22
SD-NP ($T_{\text{max}} = 1$)	0.15	**0.05**	0.05	0.06	0.10	0.12	0.09	0.15
SD-NP ($T_{\text{max}} = 2$)	**0.10**	0.06	**0.04**	0.06	0.08	0.14	0.09	**0.10**
SD-NP ($T_{\text{max}} = 3$)	0.11	0.06	0.06	**0.05**	**0.08**	**0.10**	**0.08**	0.13
SLD-Weak	0.12	0.07	0.06	0.06	0.13	0.15	0.10	0.47
SLD-Weak + POSI	**0.07**	**0.04**	**0.04**	**0.06**	0.06	0.16	0.07	0.13
SLD-Weak ($T_{\text{max}} = 1$)	0.15	0.07	0.05	0.03	0.07	0.09	0.07	0.09
SLD-Weak ($T_{\text{max}} = 2$)	0.13	0.06	0.04	0.02	0.06	0.07	0.06	0.07
SLD-Weak ($T_{\text{max}} = 3$)	0.11	0.07	**0.04**	**0.02**	**0.06**	**0.06**	**0.06**	**0.05**
SLD-Medium	0.07	0.06	0.04	0.04	0.12	0.13	0.08	0.35
SLD-Medium + POSI	**0.06**	**0.03**	0.03	0.06	**0.06**	0.15	0.07	0.10
SLD-Medium ($T_{\text{max}} = 1$)	0.10	0.07	0.04	0.04	0.08	0.09	0.07	0.08
SLD-Medium ($T_{\text{max}} = 2$)	0.09	0.05	**0.02**	0.01	0.07	**0.07**	**0.05**	**0.04**
SLD-Medium ($T_{\text{max}} = 3$)	0.08	0.06	0.04	**0.01**	**0.06**	0.10	0.06	0.05
SLD-Strong	0.07	0.05	0.02	0.03	0.09	0.12	0.06	0.26
SLD-Strong + POSI	0.05	**0.04**	0.02	0.07	0.07	0.14	0.07	0.08
SLD-Strong ($T_{\text{max}} = 1$)	0.07	0.06	0.01	0.02	0.08	0.09	0.06	0.03
SLD-Strong ($T_{\text{max}} = 2$)	0.06	0.06	0.02	0.03	0.07	**0.07**	0.06	0.04
SLD-Strong ($T_{\text{max}} = 3$)	**0.06**	0.06	**0.01**	**0.02**	**0.06**	0.08	**0.05**	**0.03**
SLD-Max	**0.05**	0.06	**0.02**	0.05	0.07	0.10	0.06	0.18
SLD-Max + POSI	0.06	0.05	**0.02**	0.08	0.05	0.10	0.06	0.10
SLD-Max ($T_{\text{max}} = 1$)	0.06	0.05	0.04	0.02	0.07	0.08	0.05	0.03
SLD-Max ($T_{\text{max}} = 2$)	0.06	0.04	0.03	0.02	0.06	0.07	0.05	0.05
SLD-Max ($T_{\text{max}} = 3$)	0.06	**0.04**	0.03	**0.02**	**0.05**	**0.07**	**0.05**	**0.03**

Table 10: Inappropriate probability (IP) by MHSC on SD v2.1