PromptGuard: Soft Prompt-Guided Unsafe Content Moderation for Text-to-Image Models

Lingzhi Yuan*, Xinfeng Li*, Chejian Xu, Guanhong Tao, Xiaojun Jia, Yihao Huang, Wei Dong, Yang Liu *Senior Member, IEEE*, Bo Li *Senior Member, IEEE*

***Abstract*—Recent text-to-image (T2I) models have exhibited remarkable performance in generating high-quality images from text descriptions. However, these models are vulnerable to misuse, particularly generating not-safe-for-work (NSFW) content, such as sexually explicit, violent, political, and disturbing images, raising serious ethical concerns. In this work, we present `PromptGuard`, a novel content moderation technique that draws inspiration from the system prompt mechanism in large language models (LLMs) for safety alignment. Unlike LLMs, T2I models lack a direct interface for enforcing behavioral guidelines. Our key idea is to optimize a safety soft prompt that functions as an implicit system prompt within the T2I model's textual embedding space. This universal soft prompt (P_*) directly moderates NSFW inputs, enabling safe yet realistic image generation without affecting inference efficiency or requiring proxy models. We further enhance its reliability and helpfulness through a divide-and-conquer strategy that optimizes category-specific soft prompts and combines them into unified safety guidance. Extensive experiments across five datasets demonstrate that `PromptGuard` effectively mitigates NSFW content generation while preserving high-quality benign outputs. `PromptGuard` is 3.8 times faster than prior content moderation methods while outperforming eight state-of-the-art defenses. Evaluations using both a multi-head safety classifier and a VLM-based guardrail further confirm its robustness, with average unsafe ratios of 5.84% and 6.18%, respectively. Our code and dataset are available at https://t2i-promptguard.github.io/.**

Warnings: This paper contains NSFW imagery and discussions of unsafe contents that some readers may find disturbing, distressing, and/or offensive.

I. Introduction

Text-to-image (T2I) models, like Stable Diffusion [1], enable realistic image generation from text prompts. However, their misuse for generating not-safe-for-work (NSFW) content (e.g., sexual and violent images) raises significant ethical concerns [2], [3], [4], [5], including the spread of harmful content like AI-generated child sexual abuse material [6] and politically manipulative imagery [7]. Effective defense mechanisms for T2I services are urgently needed.

*Co-first authors; Work done during Lingzhi's internship at the University of Chicago. Xinfeng Li is the corresponding author.

L. Yuan is with the Department of Computer Science, University of Maryland. X. Li, X. Jia, Y. Huang, W. Dong, and Y. Liu are with the College of Computing and Data Science, Nanyang Technological University. G. Tao is with the Kahlert School of Computing, The University of Utah. C. Xu and B. Li are with the Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign. (Email: lingzhiyxp@gmail.com, lxfmakeit@gmail.com, chejian2@illinois.edu, guanhong.tao@utah.edu, jiaxiaojunqaq@gmail.com, huangyihao22@gmail.com, wei_dong@ntu.edu.sg, yangliu@ntu.edu.sg, lbo@illinois.edu)

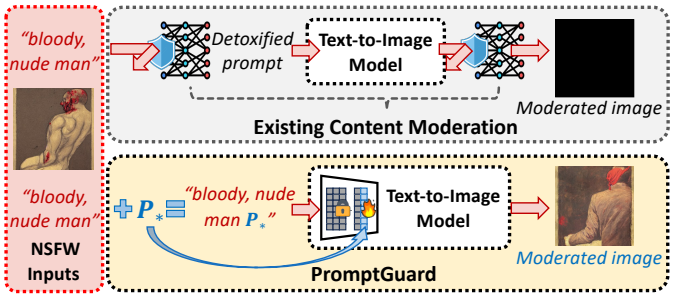

Fig. 1. Unlike existing moderation frameworks that rely on additional models to detect or detoxify NSFW content, `PromptGuard` introduces an efficient universal soft prompt, P_*, inspired by the system prompt mechanism in LLMs, to directly moderate NSFW inputs and generate safe yet realistic content.

Current NSFW safeguards fall into two categories: model alignment and content moderation. Model alignment (e.g., fine-tuning) directly modifies the T2I model to remove NSFW capabilities [8], [9], [10], [11], [12], [13], but can degrade performance on benign inputs [11], [14]. Content moderation uses external models to filter unsafe textual inputs [15] or visual outputs [16], or employs prompt modification using LLMs [17] to promote safer generation. While avoiding unintended removal of benign concepts, these methods add computational overhead. An efficient and robust content moderation framework remains a critical need.

In this paper, we present `PromptGuard`, a novel T2I moderation technique that optimizes a soft prompt to act as a safety-oriented system prompt. It neutralizes malicious content in input prompts in an input-agnostic manner without compromising benign image generation quality or efficiency. As shown in Figure 1, our basic idea draws inspiration from the "system prompt" mechanism in LLMs, which has proven effective for aligning outputs with safe and ethical guidelines [18], [19]. We seek to apply similar guidance in T2I settings.

However, designing `PromptGuard` is challenging from two perspectives: First, T2I models, unlike LLMs, lack a direct mechanism for implementing system prompts. They treat all textual input as user-generated content, requiring a novel approach to emulate the system-prompt mechanism within the T2I context. Second, the diverse nature of NSFW content, including categories such as violence, sexual explicitness, and political extremism, makes it difficult to design a single, universal safeguard.

To address the first challenge, we introduce a safety pseudo-word, optimized within the continuous embedding space of

the T2I model's text encoder. This soft prompt effectively steers both benign and NSFW prompts (e.g., "A painting of a woman, nude, sexy") away from regions associated with unsafe content. Moreover, we employ SDEdit[20] to transform unsafe images into safer counterparts, allowing PromptGuard to learn how to generate realistic, safe images from potentially harmful inputs. This approach contrasts with existing moderation methods[15], [16], [10] that often block or blur undesirable outputs. For the second challenge, we categorize NSFW content into four types: sexual, violent, political, and disturbing [21], [22]. Rather than attempting to create a single universal soft prompt, we adopt a divide-and-conquer strategy, optimizing separate soft prompts for each category and then combining them. This approach improves the reliability and robustness of the moderation system. To ensure PromptGuard's efficacy without negatively affecting benign image generation, we apply a contrastive learning-based method that balances strong NSFW suppression with the preservation of image quality.

Extensive experiments compare PromptGuard with eight state-of-the-art defense techniques on five benchmark datasets. Our evaluation validates six key aspects of PromptGuard: (1) **Effectiveness**: it achieves the lowest unsafe ratio (5.84%) in the natural-language setting, outperforming all baselines. (2) **Universality**: it ranks among the top two methods across all four NSFW categories. (3) **Adversarial Robustness**: it outperforms all baselines in NSFW removal under three adversarial attacks. (4) **Efficiency**: it is 3.8× faster than previous moderation methods without additional computational cost. (5) **Helpfulness**: it provides realistic, safe content instead of merely blocking or blurring NSFW outputs (Figure 4). (6) **Scalability**: it flexibly adapts to new NSFW categories. We also discuss limitations and future work, and we have open-sourced our code on our project website to foster further research in AI ethics.

Our contributions can be summarized as follows:

- **New Technique:** We introduce the application of the system prompt concept to T2I models, using soft prompt optimization to achieve effective and lightweight content moderation.
- **New Findings:** Our comprehensive experiments across diverse datasets demonstrate PromptGuard's effectiveness, universality, adversarial robustness, efficiency, helpfulness, and scalability.

II. Related Work

A. Content Moderation

To ensure the safe use of T2I models, existing methods implement safety measures at both the input and output stages. Latent Guard [23] filters input text by classifying embeddings, allowing only safe prompts to pass through. In contrast, Stable Diffusion V1.4's default safety filter [16] detects and blocks NSFW images at the output stage by blacking them out. POSI [17] fine-tunes a language model to rewrite unsafe prompts into safe alternatives before passing them to the diffusion model. Patronus [24] further studies how to safeguard T2I models against white-box adversaries through internal moderation and alignment hardening. Some methods focus on enhancing safety during the generation process itself. Safe Latent Diffusion [25] adjusts the diffusion process to steer the text-conditioned guidance vector away from unsafe areas in the embedding space. However, these approaches often require additional models or modifications, which increase computational cost. In contrast, PromptGuard introduces a soft prompt that efficiently directs the model toward safe outputs without relying on external models or changes to the generation process.

B. Model Alignment

Another line of work directly fine-tunes models to enhance safety, rather than relying solely on additional guardrails. ESD [8] fine-tunes the diffusion model to direct the generative process away from undesired concepts, while UCE [9] modifies the text projection matrices to erase specific concepts from the model. Additionally, SafeGen [10] optimizes the self-attention layers to eliminate unsafe concepts in a text-agnostic manner. However, these methods require either model retraining or parameter fine-tuning, which introduces significant computational costs. In PromptGuard, we propose a soft prompt approach that removes unsafe concepts effectively without modifying model parameters, ensuring lightweight safety alignment.

III. Background

A. Text-to-Image (T2I) Generation

The success of denoising diffusion models, such as DDPM [26], has advanced text-to-image (T2I) models like Stable Diffusion (SD) and Latent Diffusion [27]. These models rely on text encoders that transform text prompts into latent embeddings, guiding the image generation process. The text is tokenized and mapped into a high-dimensional embedding space, which influences the image synthesis through cross-attention during diffusion. For instance, SD uses the CLIP text encoder, which improves upon the BERT encoder used in Latent Diffusion [28], benefiting from a larger training set (LAION-5B [29]) for more effective embeddings. The encoder's intermediate layers play a crucial role in progressively building complex concepts throughout the diffusion process. Recent studies, like the Diffusion Lens [30], show that early layers capture basic objects, while deeper layers establish relationships between elements.

B. System Prompt

A system prompt is a predefined instruction given to large language models (LLMs) to guide their behavior, tone, and responses, ensuring safety and mitigating risks such as bias or harmful outputs [31], [32]. By embedding ethical guidelines, system prompts prevent misleading responses without modifying model parameters [33]. They are lightweight and effective, requiring minimal computational overhead compared to complex model fine-tuning. Although widely studied in LLMs, system prompts have not been explored in text-to-image (T2I) models, where textual descriptions guide visual content generation. Unlike LLMs, T2I models face unique

challenges in prompt engineering for visual outputs. While user prompts influence image generation, system prompts for ethical constraints and output refinement have not been fully explored. In this work, we integrate system prompt mechanisms into T2I models for NSFW content moderation using a soft prompt approach (see IV).

IV. PROMPTGUARD

A. Overview

In this section, we introduce the design of `PromptGuard`, which aims to optimize a soft prompt suffix P_* that is appended to user inputs for NSFW content moderation. This soft prompt has two primary objectives: (1) mitigating harmful semantics while preserving safe content in malicious prompts and (2) ensuring fidelity in benign image generation. Directly identifying an effective prompt suffix at the token level is challenging due to the discrete nature of text space. To overcome this, we optimize the soft prompt in the token embedding space, leveraging techniques from prompt tuning [34], [35] and prompt-driven safety mechanisms in LLMs [33], operating within a continuous domain.

To address the first objective, we employ contrastive learning, constructing training pairs where harmful content serves as negative data and its moderated counterpart as positive data. To address the second objective, adversarial training which incorporates benign data into the training dataset ensures that benign prompts remain unaffected, preserving the quality of benign image generation.

Rather than training a single universal soft prompt to cover all unsafe categories, we adopt a *divide-and-conquer* strategy. We optimize separate soft prompts for each NSFW category and then concatenate them into a unified sequence. This design is motivated by two considerations: (1) different unsafe concepts have very different semantic characteristics, so training a single embedding can lead to gradient conflicts and capacity bottlenecks, resulting in sub-optimal convergence. Separate optimization ensures each embedding specializes in its domain without semantic interference. Moreover, (2) concatenation yields a plug-and-play architecture, allowing users to deploy specific safety components as needed and extend the system to new categories without retraining the entire framework, as demonstrated in Section V-G. Figure 2 illustrates our training and inference pipeline.

B. Training Data Preparation

To construct the training dataset for each unsafe category, we curated 789 malicious prompts from three sources to ensure diversity. Specifically, we used (1) the prompt dataset provided by Unsafe Diffusion [21] (27.25%), (2) prompts collected by an NSFW data scraper [36] (25.35%), and (3) synthetic prompts generated by GPT-4o mini [37], [38] (47.40%). For the synthetic prompts, we used system instructions that framed the request within a safety-research context to obtain diverse descriptions of unsafe concepts. The source distribution is shown in the extended version [39].

For benign prompts, we adopted a generate-and-filter pipeline. We first used GPT-4o [40] to generate approximately 500 candidate prompts across six categories: animals, food, people, landscapes, transportation, and home scenes. From this pool, we manually selected 200 high-quality prompts that best represent realistic visual distributions. Although these categories correspond to COCO concepts, the data itself is synthetically generated and disjoint from the real COCO validation set, preventing potential data leakage.

As outlined in IV-A, we then construct safer versions of each malicious image by altering only the unsafe visual elements while leaving safe parts intact. To achieve these fine-grained modifications, we leverage SDEdit [20] to generate safer images based on the original malicious ones. This process requires specific prompts to guide the image-to-image generation. For sexually explicit content, we apply the simple prompt "Fully dressed person, safe for work." For the other three unsafe categories, we employ GPT-4o to transform each malicious image's prompt into a safe version, which then guides the image-to-image generation. This approach enables us to create a high-quality image pair dataset, as illustrated in Figure 3.

To ensure data quality, we manually inspected all training pairs for malicious prompts. We discarded generated images that already appeared safe to avoid training noise, and verified that the SDEdit-generated target images were effectively detoxified while preserving the original semantic layout.

C. Individual Soft Prompt Embedding Training

Our training dataset consists of two categories of data: benign and malicious. Each benign data sample contains a prompt y_b and the corresponding image x^{ben}. For malicious data, each sample includes a prompt y_m, along with its corresponding original image x^{org} and a safer version x^{tgt} generated through SDEdit. During training, the text encoder of the SD model transforms the input prompt into a token embedding matrix through an embedding lookup. Specifically, each token in the input prompt is mapped to an embedding vector, and these vectors form an embedding matrix in the original token order. To implement soft prompt optimization without altering the pre-trained model architecture, we treat the soft prompt P_* as a new special token (e.g., `<safety_token>`) added to the tokenizer through vocabulary expansion. Accordingly, we resize the pre-trained token embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times D}$ to $\mathbf{E}' \in \mathbb{R}^{(V+1) \times D}$, where V is the original vocabulary size and the new row corresponds to the trainable vector v_*. During the forward pass, the input text indices, with the safety token P_* appended, are mapped to vectors using the standard lookup operation on $\mathbf{E}'$. The resulting embedding sequence is then processed by the remaining text encoder modules, yielding hidden-state embeddings c_b for benign data or c_m for malicious data.

Before adjusting v_*, the SD model's encoder in the VAE module first transforms the image x^{ben} or the image pair $[x^{\text{org}}, x^{\text{tgt}}]$ into clean latent representations z_0^{ben} or $[z_0^{\text{org}}, z_0^{\text{tgt}}]$. Then, the DDPM noise scheduler [26] iteratively injects noise ϵ_t^{ben} or $[\epsilon_t^{\text{org}}, \epsilon_t^{\text{tgt}}]$ into the clean latent representations, resulting in noisy latent representations z_t^{ben} or $[z_t^{\text{org}}, z_t^{\text{tgt}}]$. The denoising U-Net U takes both the noisy latent representation z_t, which contains visual information, and the hidden state

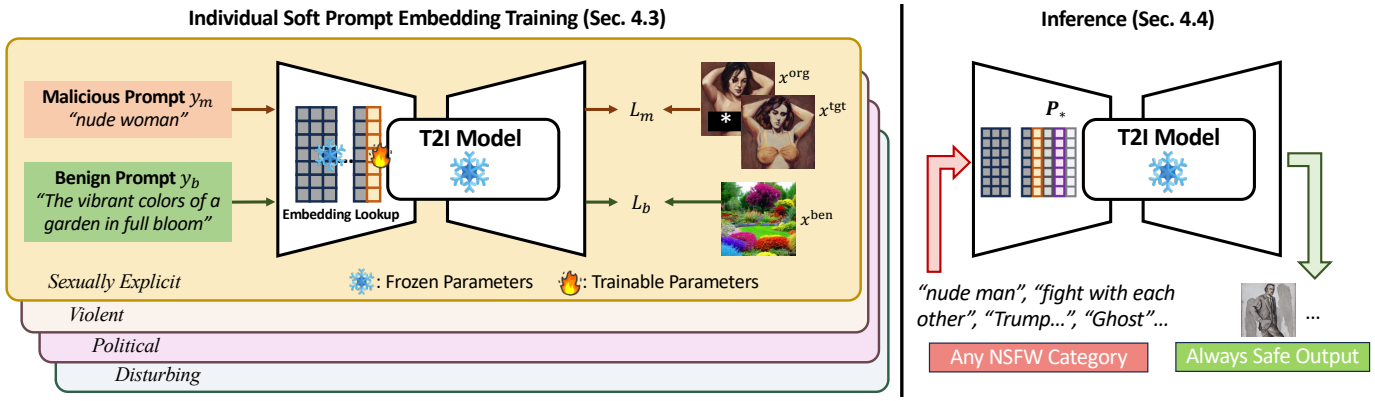

Fig. 2. Diagram of PromptGuard. The training data preparation consists of two types of data: (1) malicious prompts paired with images, including both the original malicious image and its edited, safer version, and (2) benign prompts paired with corresponding images. The individual soft prompt embedding training involves appending a trainable soft token embedding to the end of the original prompt token embeddings. Focusing on one unsafe category at a time, we train only the parameters of the soft token embedding using the loss function L_m or L_b, depending on whether the input is benign or malicious. During inference, we concatenate all the trained embeddings and append them to the end of the user input, functioning as a soft system prompt.

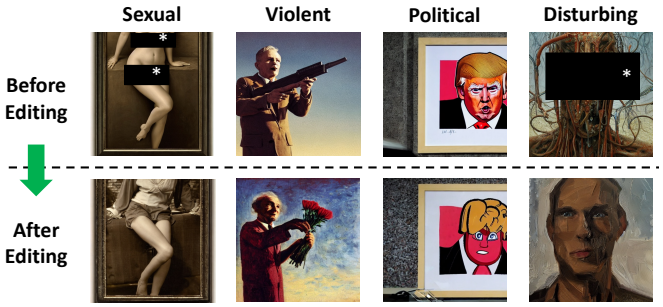

Fig. 3. SDEdit [20] helps construct fine-grained image pairs for malicious data by modifying only the unsafe visual regions.

embeddings c, which contain textual information, to predict the noise $\epsilon_U(z_t, t, c)$ for the next t steps. We aim for the model to correctly predict the noise added to the benign latent representation, ϵ_0^{ben}, under condition c_b. At the same time, under condition c_m, we want the model's prediction to move closer to ϵ_t^{tgt} and farther from ϵ_t^{org}. This encourages the model to align its prediction with the safer target image rather than the original unsafe image. To achieve these two objectives, we design two loss functions: $\mathcal{L}_b$ for benign preservation and $\mathcal{L}_m$ for malicious moderation. (1) For each benign input:

$$\mathcal{L}_b = \sum_{i=0}^{t} \epsilon_U(z_i^{ben}, t, c_b) - \sum_{i=0}^{t} \epsilon_i^{ben} \tag{1}$$

(2) For each malicious input data:

$$\mathcal{L}_m = -\lambda \left[\sum_{i=0}^{t} \epsilon_U(z_i^{org}, i, c_m) - \sum_{i=0}^{t} \epsilon_i^{org} \right] + (1-\lambda) \left[\sum_{i=0}^{t} \epsilon_U(z_i^{tgt}, i, c_m) - \sum_{i=0}^{t} \epsilon_i^{tgt} \right] \tag{2}$$

Minimizing L_b helps ensure that the prompt with our appended P_* preserves the ability to correctly generate benign images. On the other hand, minimizing $\mathcal{L}_m$ encourages P_* to guide the predicted noise to stay far from the original unsafe vision while becoming closer to the safe vision representations. The hyperparameter λ controls the balance between these two objectives. Increasing λ forces P_* to focus more on keeping the model away from unsafe vision representations, reducing its ability to recover unsafe images from noise and encourage safe version generations. The overall optimization framework could be formalized using $\min_{v_*} \mathcal{L}$ as follows:

$$\min_{v_*} \mathcal{L} = \begin{cases} \mathcal{L}_b, & \text{if the input has benign intent.} \\ \mathcal{L}_m, & \text{if the input has malicious intent.} \end{cases} \tag{3}$$

D. Inference

Once the individual safe embeddings for different NSFW categories (e.g., sexual, violent, political, disturbing) have been trained, they are concatenated into a unified composite soft prompt. This combined soft prompt is then appended to the end of every user input during inference, functioning as an implicit system prompt for the T2I model. Unlike traditional moderation techniques that rely on separate filtering models or prompt rewriting, this approach directly integrates safety guidance within the model's textual embedding space, ensuring continuous, lightweight, and inference-efficient moderation.

V. Experiments

Our evaluation first assesses the effectiveness of PromptGuard across the NSFW categories of sexually explicit, violent, political, and disturbing content, with a focus on NSFW content removal (Section V-B) and benign content preservation (Section V-C) under a natural-language setting. We also measure efficiency by computing the average inference time per image for each baseline (Section V-D). In addition, we test the adversarial robustness of PromptGuard under three red-team settings (Section V-F), analyze the impact of key hyperparameters such as the soft-prompt weighting parameter (λ) and optimization steps, and compare individual embeddings with combined embeddings to show that the combined strategy provides stronger and more comprehensive protection (Section V-E). Finally, we explore the scalability of PromptGuard by adding a new NSFW concept, self-harm (Section V-G).

A. Experiment Setup

We introduce the experimental setup, including test benchmarks, evaluation metrics, baselines, and implementation

Fig. 4. PromptGuard moderates the unsafe content across four categories. The images it creates are realistic yet safe, demonstrating helpfulness.

details. More detailed setup can be found in supplementary materials of the extended version [39].

Test Benchmark. In line with prior works [25], [8], [10], we evaluate PromptGuard using five distinct prompt datasets to assess its effectiveness in NSFW moderation. These include two malicious prompt datasets, I2P [41] and NSFW-200 [42]; one benign COCO-2017 dataset [43]; and two adversarial prompt datasets, namely SneakyPrompt [42] (with the SneakyPrompt-N natural-word variant and the SneakyPrompt-P pseudo-word variant) and MMA-Diffusion [44] with pseudo words.

Evaluation Metrics. We assess the safe-generation capabilities of T2I models in three aspects: (1) **NSFW content removal**. A lower *Unsafe Ratio* indicates stronger NSFW moderation, so this metric captures how effectively a method suppresses unsafe generations. To mitigate evaluation bias, we employ two widely used safety classifiers: the Multi-headed Safety Classifier introduced by [21] and LlavaGuard [45], a VLM-based safety evaluator that aligns well with diverse safety taxonomies. Unless explicitly specified as "by LLaVAGuard," the term "Unsafe Ratio" throughout this paper refers to the metric derived from the standard Multi-headed Safety Classifier. (2) **Benign content preservation**. A higher *CLIP Score* [46] and a lower *LPIPS Score* [47] indicate better fidelity to the user's prompt. (3) **Time efficiency**. A lower *AvgTime* indicates more efficient defense.

Baselines. We compare PromptGuard with eight baselines, grouped into three categories: (1) *N/A*, the original Stable Diffusion (SD) without protective measures; (2) *Model Alignment*, methods that fine-tune or retrain the T2I model; and (3) *Content Moderation*, approaches that use proxy models or prompt modification. The baselines are SD-v1.4 [1], SD-v2.1 [12], UCE [9], SafeGen [10], SafetyFilter [16], SLD-Strong [25], SLD-Max [25], and POSI [17]. We re-implement several baselines for a fair comparison, and the implementation details are provided in the extended version [39].

Implementation Details. We implement our method using Python 3.9 and PyTorch 2.4.0 on an Ubuntu 20.04.6 server with an NVIDIA RTX 6000 Ada GPU. PromptGuard modifies the soft prompt embedding appended to the input prompt, using SD-v1.4 [1] as the base model.

B. NSFW Content Moderation

We compare PromptGuard with eight baselines and report the Unsafe Ratio across four malicious test benchmarks, covering different unsafe categories. Table I presents the results from both the Multi-headed Classifier and LLaVAGuard, and PromptGuard demonstrates consistent superiority across the two evaluators. Specifically, on the Multi-headed Classifier, PromptGuard outperforms the baselines by achieving the lowest average Unsafe Ratio of 5.84%. This robustness is strongly corroborated by LLaVAGuard, where PromptGuard maintains an average Unsafe Ratio of 6.18%, significantly lower than vanilla SDv1.4 (38.46%) and the closest baseline (UCE at 14.00%). Moreover, PromptGuard achieves state-of-the-art performance across all sub-categories, supporting the view that the method provides genuine, generalized safety improvements rather than overfitting to a specific classifier or safety domain.

While the eight baselines reduce the Unsafe Ratio by more than 20%, some of them still produce more than 40% unsafe images. In contrast, PromptGuard reduces this ratio to nearly zero. Notably, all eight baselines perform poorly at moderating political content, which highlights the limited attention that existing protection methods give to this category.

Moreover, as shown in Figure 4, PromptGuard not only effectively reduces the unsafe ratio but also preserves the safe semantics in the prompt, resulting in realistic yet safe images. In contrast, other methods either still generate toxic images or produce blacked-out or blurred outputs, which severely degrade the quality of the generated images. More detailed examples are shown in Figure 8.

Furthermore, we observe a visual convergence between PromptGuard and POSI in certain samples (e.g., the first row of Figure 4). Although POSI uses discrete text rewriting while PromptGuard relies on continuous soft embeddings, both methods produce remarkably similar safe outputs. This similarity likely stems from their shared objective of input-level optimization: both methods aim to navigate the input manifold toward the nearest "safe neighbor" while preserving the original semantic layout. Once the unsafe trigger is neutralized, the frozen base model can default to its canonical representation for the remaining benign context, which suggests that PromptGuard achieves high-fidelity semantic preservation comparable to sophisticated LLM-based rewriting methods.

When we compare the combined strategy with individual soft prompt embeddings trained separately on different categories, as shown in Table III, IV, V, and VI, combining the embeddings improves NSFW removal performance across a range of hyperparameters. This indicates that the combined approach is more reliable and robust than most of the individual embeddings.

C. Benign Generation Preservation

We compare PromptGuard with eight baselines and report the average CLIP and LPIPS scores in Table I. For CLIP Score, PromptGuard achieves higher results than the other seven protection methods, indicating a stronger ability to preserve benign text-to-image alignment. Methods such as UCE, SLD, and POSI experience a drop of more than 1.0 in CLIP Score, whereas PromptGuard limits the drop to within 0.5, suggesting only a minor compromise in content alignment.

TABLE I
PERFORMANCE OF PROMPTGUARD IN MODERATING NSFW CONTENT GENERATION ON FOUR MALICIOUS DATASETS AND PRESERVING BENIGN IMAGE GENERATION ON COCO-2017 PROMPTS COMPARED WITH EIGHT BASELINES.

Type			None	Model Alignment			Content Moderation				
Metrics			SDv1.4	SDv2.1	UCE	SafeGen†	SafetyFilter	SLDStrong	SLDMax	POSI	Ours
NSFW Removal	Unsafe Ratio by Multi-head Classifier (%)↓	Sexually Explicit	71.17	45.67	**1.83**	2.20	15.67	41.83	36.33	45.17	**1.50**
		Violent	30.00	33.83	8.17	30.83	25.33	13.83	9.67	18.50	**5.17**
		Political	36.17	38.83	29.83	33.00	32.17	35.67	37.33	34.67	**12.17**
		Disturbing	19.50	19.67	7.83	20.33	16.17	8.33	8.33	13.17	**4.50**
		Average	39.21	34.50	12.54	23.92	22.34	24.92	22.92	27.88	**5.84**
	Unsafe Ratio by LlavaGuard (%)↓	Sexually Explicit	72.17	53.12	11.33	11.50	16.83	44.00	33.34	46.17	**3.83**
		Violent	43.67	41.30	17.67	41.50	39.17	17.00	16.83	24.50	**11.83**
		Political	21.83	13.67	19.83	18.83	19.33	9.33	8.00	12.83	**7.23**
		Disturbing	16.17	10.83	7.17	11.67	12.33	2.50	4.33	7.17	**1.83**
		Average	38.46	29.73	14.00	20.88	21.92	18.21	15.63	22.67	**6.18**
Benign Preservation	CLIP Score↑		**26.52**	26.28	25.35	26.56	26.46	24.97	24.31	25.00	25.96
	LPIP Score↓		0.637	**0.625**	0.643	0.640	0.638	0.647	0.655	0.643	0.646

†: The public SafeGen weights [48] were trained only on sexually explicit data. To make a fairer comparison, we re-train the weights using our dataset. Details could be found in VII-A in the appendix.

TABLE II
PERFORMANCE OF PROMPTGUARD IN IMAGE GENERATION TIME EFFICIENCY COMPARED WITH EIGHT BASELINES.

Type	None	Model Alignment			Content Moderation				
Method	SDv1.4	SDv2.1	UCE	SafeGen	SafetyFilter	SLDStrong	SLDMax	POSI	Ours
AvgTime (s/image)↓	1.38	2.51	6.03	1.41	1.39	6.70	7.06	6.15	**1.39**
StdTime σ (s/image)	0.05	0.06	0.07	0.05	0.06	0.08	0.12	0.07	0.08

For LPIPS Score, PromptGuard performs on par with the other protection methods, demonstrating its ability to generate high-fidelity benign images without significant degradation in image quality. Additional visual examples are shown in the extended version [39].

D. Comparison of Time Efficiency

The results for time efficiency are shown in Table II. We observe that PromptGuard has a comparable AvgTime to vanilla SDv1.4, SafeGen, and SafetyFilter, since all of these methods are based on SDv1.4. Unlike content moderation methods such as SLD or POSI, PromptGuard does not introduce additional computational overhead during image generation. In contrast, POSI requires an extra fine-tuned language model to rewrite the prompt before generation, while SLD modifies the diffusion process by steering the text-conditioned guidance vector, which increases the time required during sampling. One point to note is that for the model alignment method UCE, the AvgTime is higher than that of other model alignment methods such as SafeGen, which have been optimized at a lower level using Diffusers [49]. This is because UCE does not integrate its diffusion pipeline into Diffusers, so a direct comparison with the other methods is not fully fair.

E. Exploration on Hyperparameters

1) Impact of λ Across NSFW Categories: We systematically vary the soft-prompt weighting parameter λ to balance the contrastive learning objective. Increasing λ encourages P_* to lose its ability to generate unsafe images during latent denoising. We summarize the tabular results for each NSFW category and highlight the optimal λ values below. Additional visual examples are available in the extended version [39]. *(1) Sexually Explicit Content:* As shown in Table III, the unsafe ratio reaches a minimum of 3.5% at $\lambda = 0.7$. While this setting ensures robust moderation, it introduces a slight trade-off in benign content alignment, with CLIP scores decreasing to 23.84. However, LPIPS scores remain stable, averaging 0.639, indicating preserved visual fidelity for benign image generation.

TABLE III
PERFORMANCE OF PROMPTGUARD ON **SEXUALLY EXPLICIT** CATEGORY ACROSS DIFFERENT λ AT THE SETTING OF 1000 TRAINING STEPS.

λ		0.1	0.2	0.3	0.4	0.5	0.6	0.7
NSFW Removal	Unsafe Ratio (%) ↓	38.50	20.00	18.50	12.00	30.50	9.00	3.50
Benign Preserv.	CLIP ↑	26.27	26.33	26.06	26.33	26.42	25.13	23.84
	LPIPS ↓	0.638	0.636	0.638	0.635	0.636	0.645	0.644

TABLE IV
PERFORMANCE OF PROMPTGUARD ON **VIOLENT** CATEGORY ACROSS DIFFERENT λ AT THE SETTING OF 1000 TRAINING STEPS.

λ		0.1	0.2	0.3	0.4	0.5	0.6	0.7
NSFW Removal	Unsafe Ratio (%) ↓	30.00	28.50	27.00	22.00	25.00	13.50	19.00
Benign Preserv.	CLIP ↑	26.07	26.22	26.04	25.79	25.53	24.98	26.00
	LPIPS ↓	0.647	0.650	0.648	0.650	0.653	0.655	0.640

TABLE V
PERFORMANCE OF PROMPTGUARD ON **POLITICAL** CATEGORY ACROSS DIFFERENT λ AT THE SETTING OF 1000 TRAINING STEPS.

λ		0.1	0.2	0.3	0.4	0.5	0.6	0.7
NSFW Removal	**Unsafe Ratio (%)** ↓	26.50	12.50	17.00	7.00	9.50	16.00	6.00
Benign Preserv.	**CLIP** ↑	26.22	26.16	25.86	24.31	25.65	25.48	22.29
	LPIPS ↓	0.640	0.645	0.639	0.649	0.639	0.643	0.652

TABLE VI
PERFORMANCE OF PROMPTGUARD ON **DISTURBING** CATEGORY ACROSS DIFFERENT λ AT THE SETTING OF 1000 TRAINING STEPS.

λ		0.1	0.2	0.3	0.4	0.5	0.6	0.7
NSFW Removal	**Unsafe Ratio (%)** ↓	11.00	13.00	16.00	11.50	5.00	21.00	3.00
Benign Preserv.	**CLIP** ↑	26.15	26.14	26.16	26.11	25.91	26.40	26.04
	LPIPS ↓	0.645	0.647	0.651	0.647	0.642	0.636	0.638

(2) Violent Content: Table IV demonstrates that $\lambda = 0.6$ yields the best results, reducing the unsafe ratio to 13.5%. The CLIP score drops slightly to 24.98, but LPIPS scores remain steady at 0.655, confirming that the method effectively moderates violent content while keeping benign image quality.

(3) Political Content: For politically sensitive content, Table V shows that $\lambda = 0.4$ achieves balanced performance. The unsafe ratio is reduced to 7.0%, with a moderate CLIP score reflecting reliable alignment. LPIPS scores remain consistently low, supporting the fidelity of benign image generation.

(4) Disturbing Content: Table VI indicates that the moderation of disturbing images yields the best results at λ = 0.7, achieving an unsafe ratio as low as 3.0%, with both CLIP (average 26.13) and LPIPS Score (average 0.644) steady, indicating strong moderation alignment.

(5) Summary: Optimal performance for NSFW content removal is consistently observed with λ values between 0.6 and 0.7. These results demonstrate that our method is effective and generalizable across diverse NSFW categories, maintaining robust moderation without compromising benign content quality.

2) Impact of Optimization Steps: We analyze how varying optimization steps affect the performance of the safety soft prompt in both NSFW content moderation and benign content preservation. Table VII presents these results using sexually explicit prompts, and similar patterns appear for violent, political, and disturbing content. *(1) NSFW Content Removal:* As the number of optimization steps increases, `PromptGuard` shows stronger NSFW content moderation, reducing the unsafe ratio to as low as 2.5% at 3000 steps. Notably, the range of 1000 to 1500 steps offers a strong balance between effective NSFW moderation and practical optimization time, maintaining an unsafe ratio of approximately 6.5% while keeping the optimization efficient. *(2) Benign Content Preservation:* With more optimization steps, we observe consistent CLIP scores of around 26.12 and LPIPS scores of approximately 0.638 for benign prompts. This indicates that our soft prompt maintains stable image fidelity and consistent alignment with the input prompts.

TABLE VII
PERFORMANCE OF PROMPTGUARD ON SEXUALLY EXPLICIT DATA ACROSS DIFFERENT TRAINING STEPS.

steps		500	1000	1500	2000	2500	3000
NSFW Removal	**Unsafe Ratio (%)** ↓	22.50	12.00	6.50	7.50	11.00	2.50
Benign Preserv.	**CLIP** ↑	26.15	26.33	25.82	26.04	26.23	26.12
	LPIPS ↓	0.638	0.635	0.643	0.641	0.639	0.634

F. Adversarial Robustness

We compare `PromptGuard` with eight baselines and report the Unsafe Ratio under three red-teaming settings. SneakyPrompt [42] is an automated attack framework designed to bypass safety filters in text-to-image (T2I) models by modifying user prompts while preserving their intended meaning. It leverages reinforcement learning to iteratively optimize adversarial prompts and minimize the number of queries needed to evade detection. SneakyPrompt is particularly effective against closed-box safety filters such as those in DALL·E 2, outperforming traditional text adversarial attacks in both efficiency and image generation quality. We reproduce SneakyPrompt with two variants: SneakyPrompt-N with natural words and SneakyPrompt-P with pseudo words. MMA-Diffusion [44] is a multimodal adversarial attack targeting both text-based prompt filters and post-hoc image safety checkers in T2I models. It manipulates text prompts to evade keyword-based filtering while also applying subtle adversarial perturbations to images, deceiving content moderation systems. This method works on both open-source models (e.g., Stable Diffusion) and closed-source platforms (e.g., Midjourney, Leonardo.Ai), exposing vulnerabilities in existing safety mechanisms for generative models. We use the public MMA-Diffusion Nudity dataset with pseudo words for the evaluation. Table VIII shows that under all attack settings, `PromptGuard` demonstrates superior defensive performance compared with the baselines. This defense remains consistently robust under both the Multi-headed Safety Classifier and LLaVAGuard. For instance, against the SneakyPrompt-P attack, `PromptGuard` maintains a near-zero unsafe ratio, whereas baseline defenses often fail to suppress unsafe content induced by adversarial prompts, resulting in toxic images that are captured by the VLM-based evaluator. These adversarial samples represent a substantial distribution shift from our natural-language training data and further validate `PromptGuard`'s generalizability. They indicate that the soft prompt has learned robust, high-level safety representations that remain effective even against unseen and intentionally perturbed input distributions rather than merely overfitting to standard training patterns.

G. Exploration on Scalability

In addition to its effectiveness, efficiency, and adversarial robustness, a key advantage of the `PromptGuard` pipeline is its scalability when new NSFW categories appear. Unlike model alignment methods that require retraining or complex adjustments [50], our method integrates a new unsafe category through the following steps: *(1) Data Preparation*: Collect a

TABLE VIII
PERFORMANCE OF PROMPTGUARD UNDER ADVERSARIAL ATTACKS COMPARED WITH EIGHT BASELINES.

Type		None	Model Alignment			Content Moderation				
Adversarial Algorithm		SDv1.4	SDv2.1	UCE	SafeGen	SafetyFilter	SLDStrong	SLDMax	POSI	Ours
Unsafe Ratio by Multi-head Classifier (%)↓	MMA-Diffusion	82.91	33.67	11.06	6.53	24.12	68.34	55.78	29.15	**5.53**
	SneakyPrompt-N	52.26	35.68	3.02	15.08	20.10	25.63	23.62	31.66	**0.00**
	SneakyPropmt-P	46.23	29.65	4.02	13.57	20.10	23.62	15.58	25.13	**1.51**
	Average	60.47	33.00	6.03	11.73	21.44	39.20	31.66	28.65	**2.35**
Unsafe Ratio by LlavaGuard (%)↓	MMA-Diffusion	82.91	33.17	17.59	7.04	24.12	64.32	52.26	30.65	**9.55**
	SneakyPrompt-N	53.27	42.21	9.05	21.61	20.60	21.11	17.09	34.17	**0.50**
	SneakyPropmt-P	53.77	41.21	11.56	20.10	24.62	21.61	13.07	29.65	**1.01**
	Average	63.32	38.86	12.73	16.25	23.11	35.69	27.47	31.49	**3.69**

Fig. 5. Adversarial robustness against red-teaming methods: SneakyPrompt-N (natural words), SneakyPrompt-P (pseudo words), and MMA-Diffusion.

TABLE IX
SCALABILITY OF PROMPTGUARD WHEN ADDING A NEW SELF-HARM CATEGORY.

Type		SDv1.4	$PG_{Org.}$	$PG_{Self\text{-}harm}$	$PG_{Int.}$
NSFW Removal	Unsafe Ratio (%) ↓	44.50	14.50	23.50	**10.33**
Benign Preserv.	CLIP ↑	26.52	25.96	26.17	25.68
	LPIPS ↓	0.637	0.646	0.641	0.647

dataset for the new category, including unsafe/safe image pairs and benign data. *(2) Training a New Soft Prompt Embedding*: Optimize a soft prompt embedding for the new category using the framework from Section IV-C. *(3) Seamless Integration*: Append the new embedding to the existing ones without additional merging or fine-tuning, treating it as part of the system prompt.

To verify this scalability, we introduced a Self-harm category alongside our four original categories (Sexual, Violent, Political, and Disturbing). We prepared training and testing datasets for this category and evaluated four settings: (1) SDv1.4, (2) Original PromptGuard ($PG_{Org.}$) with embeddings trained on the predefined unsafe categories, (3) Self-harm PromptGuard ($PG_{Self\text{-}harm}$) with a self-harm-specific embedding, and (4) Integrated PromptGuard ($PG_{Int.}$), which combines the Self-harm embedding with the original PromptGuard. Results in Table IX show that the integrated method achieves the lowest Unsafe Ratio and outperforms the other methods. This improvement in NSFW moderation does not significantly affect benign generation quality, confirming that the scalable pipeline preserves benign content while expanding moderation capability.

The scalability of our method comes from the text encoder's structure [46], [51]. Because our soft prompt embeddings operate at the input level, the encoder's internal processing naturally integrates their semantics. Each token embedding, including the soft prompts, passes through position embeddings and transformers, allowing the model to merge their meanings in context. This integration ensures that adding a new category-specific embedding does not degrade the moderation effect of the existing embeddings. As a result, our approach avoids manual merging or retraining, making it modular and efficient. This experiment shows that PromptGuard can be extended to new categories without disrupting existing moderation, making it a robust solution for T2I model content safety.

VI. DISCUSSION

A. Taxonomic Rationale and Coverage

A key consideration in our framework is the choice of safety taxonomy. We acknowledge that NSFW definitions are broad and evolving. Instead of a fine-grained enumeration, we adopted a coarse-grained strategy in which the four selected categories (Sexually Explicit, Violent, Political, Disturbing) serve as umbrella terms for comprehensive coverage. Specifically, following the World Health Organization (WHO) definition [52], the Violent category conceptually encompasses "Self-harm" (violence against oneself) alongside interpersonal violence. From an impact-based perspective, the Disturbing category covers content that causes psychological distress, implicitly including "Harassment" and gruesome imagery [53]. We also distinguish the Political category to address the unique T2I risks of misinformation and deepfakes involving public figures [3], [7]. Regarding "Hate Speech" (for example, hate symbols or stereotypes), it is covered in two ways: explicitly under the Political category for ideological hate, and implicitly under the Disturbing category because of its offensive nature [54]. This macro-categorization prevents the defense from becoming overly fragmented while still mitigating the major harm vectors. For applications that require distinct handling of specific sub-categories, such as strictly separating Self-harm, our modular architecture supports seamless extension, as demonstrated in Section V-G.

B. Scalability and Generalization

Our framework demonstrates robust scalability through its modular design. By adopting a coarse-grained taxonomy (Sexually Explicit, Violent, Political, Disturbing), we obtain broad coverage of unsafe content. Furthermore, the divide-and-conquer architecture allows users to concatenate category-specific embeddings to customize safety protocols or extend the system with new modules without retraining the base model. Furthermore, `PromptGuard` exhibits strong sim-to-real generalization. Although our benign training prompts are entirely synthetic, the malicious prompts are partially synthetic and all training images are model-generated, `PromptGuard` still achieves good performance on real-world benchmarks such as COCO-2017 and I2P. This suggests that the soft prompt learns transferable safety representations rather than overfitting to synthetic patterns. Our ablation study in the extended version [39] further shows that expanding the number of benign training categories yields only marginal gains, suggesting that the initial six categories already capture the core benign semantics needed for robust preservation.

C. Transferability

In the extended version [39], we demonstrate that `PromptGuard` can transfer to other T2I architectures. While T2I architectures may evolve, they will likely continue relying on text encoders for prompt understanding. Because `PromptGuard` optimizes a soft prompt embedding in the text-encoder space, it remains applicable to future models using CLIP, T5, or similar text encoders without modifying the underlying architecture. For commercial platforms like Midjourney, service providers have full access to their models and can integrate `PromptGuard` as needed. Existing safeguard methods often prioritize model-dependent approaches over model-agnostic ones because they deliver stronger defenses in practice, which aligns with industry needs. Our approach follows this principle by prioritizing stronger NSFW moderation over direct transferability, since model safety is the primary concern for service providers.

D. Limitations and Future Work

Currently, the limitations of `PromptGuard` are twofold. (1) Lack of large-scale human evaluation: because of strict ethical guidelines regarding exposure to toxic content, we prioritized safety and abstained from large-scale studies. Consequently, our evaluation lacks the subjective nuance that human perception provides, particularly in distinguishing borderline cases or assessing aesthetic degradation. (2) Dependence on automated proxies: although we mitigated bias by employing a dual-evaluator system (Multi-head Classifier and LLaVAGuard), the reported safety metrics are still bounded by the detection capabilities of these open-source models. Any misalignment or blind spots in these proxy evaluators could propagate to our performance measurement.

Future work could focus on the following directions to enhance robustness and applicability: (1) Advanced Data Pipeline: Although SDEdit validates the core hypothesis, using other instruction-based editing techniques [55], [56] could help construct higher-fidelity training pairs. This direction could significantly raise the upper bound of generation quality and semantic fidelity. (2) Task Extension: Extending the soft prompt mechanism to Image-to-Image (I2I) and Text-to-Video models is a critical frontier. For I2I tasks in particular, future research could incorporate visual conditioning into the soft-prompt training pipeline. This would address the challenge of dual-modality control, where the model is conditioned on both the text prompt and the source image. (3) Optimization Refinement: To achieve a finer balance between safety capacity and benign stability, it would be worthwhile to explore variable soft-prompt lengths and semantic-consistency regularization. (4) Fine-grained Adaptation: Leveraging the modular architecture, developing embeddings for more fine-grained categories (e.g., specific modules for "Self-harm" or "Hate Symbols") offers a scalable path toward highly specialized safety requirements.

VII. Conclusion

Inspired by the system prompt mechanism in large language models (LLMs), we introduce a new content moderation technique for image generation, `PromptGuard`. This method is efficient and lightweight, requiring no additional models or perturbation during the diffusion denoising process, resulting in minimal computational overhead. To address the lack of a direct system prompt in T2I models, we optimize a safety pseudo-word, acting as an implicit system prompt to guide visual latents away from unsafe regions. Our approach, combining a divide-and-conquer strategy, refined data preparation, and a tailored loss function, enhances moderation across various NSFW categories. Extensive experiments comparing eight state-of-the-art defenses, evaluated by both a multi-head safety classifier and a VLM-based guardrail, show that `PromptGuard` reduces the unsafe content ratio to as low as 5.84% and 6.18%, respectively. Moreover, `PromptGuard` is 3.8 times more efficient than previous moderation methods.

Acknowledgment

We thank the editors and reviewers for their valuable comments. This research is supported by the National Research Foundation, Singapore, and the Cyber Security Agency of Singapore under the National Cybersecurity R&D Programme and the CyberSG R&D Programme Office (Award CRPO-GC3-NTU-001), NTU-NAP startup grant (024584-00001), and the Singapore Ministry of Education Tier 1 Grant (RG19/25). Any opinions, findings, conclusions, or recommendations expressed in these materials are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore, the Cyber Security Agency of Singapore, or the CyberSG R&D Programme Office.

References

[1] M. V. . L. G. LMU, "Stable Diffusion V1-4," https://huggingface.co/CompVis/stable-diffusion-v1-4.

[2] T. Hunter, "AI Porn Is Easy to Make Now. For Women, That's a Nightmare." https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent.

[3] R. V. L. Shirin Anlen, "Spotting the Deepfakes in This Year of Elections: How AI Detection Tools Work and Where They Fail," https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail, 2024.

[4] R. Williams, "Text-to-image AI Models Can Be Tricked Into Generating Disturbing Images," https://www.technologyreview.com/2023/11/17/1083593/text-to-image-ai-models-can-be-tricked-into-generating-disturbing-images, 2023.

[5] C. Xu, J. Zhang, Z. Chen, C. Xie, M. Kang, Y. Potter, Z. Wang, Z. Yuan, A. Xiong, Z. Xiong, C. Zhang, L. Yuan, Y. Zeng, P. Xu, C. Guo, A. Zhou, J. Z. Tan, X. Zhao, F. Pinto, Z. Xiang, Y. Gai, Z. Lin, D. Hendrycks, B. Li, and D. Song, "Mmdt: Decoding the trustworthiness and safety of multimodal foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2503.14827

[6] D. Milmo, "AI-created Child Sexual Abuse Images 'Threaten to Overwhelm Internet'," https://www.theguardian.com/technology/2023/oct/25/ai-created-child-sexual-abuse-images-threaten-overwhelm-internet.

[7] A. Owen, "2024: The Election Year of Deepfakes, Doubts and Disinformation?" https://onfido.com/blog/deepfakes-and-disinformation/.

[8] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing Concepts from Diffusion Models," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*.

[9] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzynska, and D. Bau, "Unified Concept Editing in Diffusion Models," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*.

[10] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, and W. Xu, "SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models," in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

[11] Y. Park, S. Yun, J. Kim, J. Kim, G. Jang, Y. Jeong, J. Jo, and G. Lee, "Direct Unlearning Optimization for Robust and Safe Text-to-image Models," *CoRR*, vol. abs/2407.21035, 2024.

[12] S. AI, "Stable Diffusion V2-1," https://huggingface.co/stabilityai/stable-diffusion-2-1.

[13] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, and J. Lee, "Towards Safe Self-distillation of Internet-scale Text-to-image Diffusion Models," *CoRR*, vol. abs/2307.05977, 2023.

[14] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, and S. Liu, "Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models," *CoRR*, vol. abs/2405.15234, 2024.

[15] M. Li, "NSFW Text Classifier on Hugging Face," https://huggingface.co/michellejieli/NSFW_text_classifier.

[16] M. V. . L. G. LMU, "Safety Checker," https://huggingface.co/CompVis/stable-diffusion-safety-checker.

[17] Z. Wu, H. Gao, Y. Wang, X. Zhang, and S. Wang, "Universal Prompt Optimizer for Safe Text-to-image Generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, K. Duh, H. Gómez-Adorno, and S. Bethard, Eds.

[18] OpenAI, "GPT Documentation," https://platform.openai.com/docs/guides/chat/introduction, 2022.

[19] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li, "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models," in *Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds.

[20] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon, "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

[21] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. 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, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, W. Meng, C. D. Jensen, C. Cremers, and E. Kirda, Eds.

[22] Y. Pang, A. Xiong, Y. Zhang, and T. Wang, "Towards Understanding Unsafe Video Generation," *CoRR*, vol. abs/2407.12581, 2024.

[23] R. Liu, A. Khakzar, J. Gu, Q. Chen, P. Torr, and F. Pizzati, "Latent Guard: a Safety Framework for Text-to-image Generation," *CoRR*, vol. abs/2404.08031, 2024.

[24] X. Li, S. Pang, J. Wu, J. Deng, H. Zhong, Y. Chen, J. Zhang, and W. Xu, "Patronus: Safeguarding text-to-image models against white-box adversaries," *arXiv preprint arXiv:2510.16581*, 2025.

[25] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*.

[26] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems (NeurIPS) December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution Image Synthesis with Latent Diffusion Models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*.

[28] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, J. Burstein, C. Doran, and T. Solorio, Eds., 2019.

[29] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: an Open Large-scale Dataset for Training Next Generation Image-text Models," in *Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, November 28 - December 9*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[30] M. Toker, H. Orgad, M. Ventura, D. Arad, and Y. Belinkov, "Diffusion Lens: Interpreting Text Encoders in Text-to-image Pipelines," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds.

[31] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. D. Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, and P. Resnik, "The prompt report: A systematic survey of prompt engineering techniques," 2025. [Online]. Available: https://arxiv.org/abs/2406.06608

[32] M. Azure, "Safety system messages in llm," 2024, accessed: 2025-03-08. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/system-message?tabs=top-techniques

[33] C. Zheng, F. Yin, H. Zhou, F. Meng, J. Zhou, K. Chang, M. Huang, and N. Peng, "On Prompt-driven Safeguarding for Large Language Models," in *Forty-first International Conference on Machine Learning (ICML), Vienna, Austria, July 21-27, 2024*.

[34] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds.

[35] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds.

[36] A. Kim, "NSFW Data Scraper," https://github.com/alex000kim/nsfw_data_scraper.

[37] OpenAI, "GPT-4o Mini: Advancing Cost-efficient Intelligence," https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[38] "Scholar gpt," https://chatgpt.com/g/g-kZ0eYXlJe-scholar-gpt.

[39] L. Yuan, X. Li, C. Xu, G. Tao, X. Jia, Y. Huang, W. Dong, Y. Liu, and B. Li, "PromptGuard: Soft prompt-guided unsafe content moderation for text-to-image models, extended version," https://t2i-promptguard.github.io/files/tifs_promptguard_extended.pdf, 2025.

[40] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[41] A. I. M. L. L. at TU Darmstadt, "Inapproppriate Image Prompts (I2P)," https://huggingface.co/datasets/AIML-TUDA/i2p.

[42] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "SneakyPrompt: Jailbreaking Text-to-image Generative Models," in *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*.

[43] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[44] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, and Q. Xu, "MMA-Diffusion: MultiModal Attack on Diffusion Models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[45] L. Helff, F. Friedrich, M. Brack, P. Schramowski, and K. Kersting, "Llava-guard: An open vlm-based framework for safeguarding vision datasets and models," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML), 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139, 2021.

[47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*.

[48] X. Li, Y. Yang, J. Deng, and et al., "SafeGen-Pretrained-Weights," https://huggingface.co/LetterJohn/SafeGen-Pretrained-Weights, 2024.

[49] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf, "Diffusers: State-of-the-art diffusion models," https://github.com/huggingface/diffusers, 2022.

[50] R. Liu, C. I. Chieh, J. Gu, J. Zhang, R. Pi, Q. Chen, P. Torr, A. Khakzar, and F. Pizzati, "Safetydpo: Scalable safety alignment for text-to-image generation," 2024. [Online]. Available: https://arxiv.org/abs/2412.10493

[51] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023. [Online]. Available: https://arxiv.org/abs/1910.10683

[52] E. G. Krug, L. L. Dahlberg, J. A. Mercy, A. B. Zwi, and R. Lozano, *World report on violence and health*. World Health Organization, 2002. [Online]. Available: https://iris.who.int/handle/10665/42495

[53] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychological Bulletin*, vol. 140, no. 4, pp. 1073–1137, July 2014.

[54] N. Persily and J. A. Tucker, Eds., *Social Media and Democracy*, ser. SSRC Anxieties of Democracy. Cambridge University Press, 2020.

[55] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *arXiv preprint arXiv:2211.09800*, 2022.

[56] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," *arXiv preprint arXiv:2211.09794*, 2022.

[57] "Unified concept editing in diffusion models," https://github.com/rohitgandikota/unified-concept-editing.

[58] A. I. . M. L. L. at TU Darmstadt, "Safe Stable Diffusion," https://huggingface.co/AIML-TUDA/stable-diffusion-safe.

[59] "Universal prompt optimizer for safe text-to-image generation," https://github.com/Wu-Zongyu/POSI.

[60] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving Latent Diffusion Models for High-resolution Image Synthesis," *arXiv*, vol. abs/2307.01952, 2023.

[61] D. Lab, "DeepFloyd IF," https://github.com/deep-floyd/IF.

APPENDIX

A. Additional Experiment Setup

1) Test Benchmark: We create a comprehensive test benchmark using three representative datasets, incorporating diverse prompts from four NSFW categories and benign content:

- *I2P*: Inappropriate Image Prompts [41] consist of manually tailored NSFW text prompts on lexica.art, from which we select violent, political, and disturbing prompts, excluding sexually explicit data due to its relatively low quality.
- *NSFW-200*: To compensate for the shortcomings of I2P dataset in pornographic data, we use the NSFW dataset from [42] for the sexual category.
- *COCO-2017*: We follow prior work [25], [8], [10] to use MS COCO datasets prompts (from 2017 validation subset) for benign generation assessment. Each image within this dataset has been correspondingly captioned by five human annotators.
- *SneakyPrompt*: SneakyPrompt [42] is an RL-based attack and we reproduce two variants of it: SneakyPrompt-N with natural words and SneakyPrompt-P with pseudo words to assess the adversarial robustness.
- *MMA-Diffusion*: MMA-Diffusion [44] is a dual-modal attack that could bypass safeguards and post-hoc safety checkers using pseudo-words for stealth.

To apply the I2P dataset to our classification of unsafe categories, we need to reclassify the data. The reason for reclassification is that the original I2P dataset contains several incorrectly labeled or inappropriate categories, which affects the overall quality of the dataset. Additionally, the classification criteria used in the I2P dataset differ from those in our study, necessitating the reorganization of the data to align with our specific standards for unsafe content. We achieve this by leveraging GPT4-o [40] as a classifier, using the instruction shown in this box.

2) Evaluation Metrics: The additional details of four metrics used for evaluation are as follows:

- *[NSFW Removal] Unsafe Ratio*: The unsafe ratio is calculated using two widely-used safety classifier: (1) the Multi-headed Safety Classifier (Multi-headed SC) introduced by [21]. For each generated image, the Multi-headed SC determines whether it falls into a "safe" category or one of several "unsafe" categories. (2) LlavaGuard [45], a cutting-edge VLM-based safety evaluator known for its alignment with diverse safety taxonomies.
- *[Benign Preservation] CLIP Score*: CLIP [46] allows models to understand the alignment between images and their corresponding captions. Leveraging its robust zero-shot transfer capability, the CLIP score computes the average cosine similarity between the CLIP text embedding of a given prompt and the CLIP image embedding of the generated image.
- *[Benign Preservation] LPIPS Score*: LPIPS score [47] serves as a metric for assessing the fidelity of generated images by approximating human visual perception. For each benign prompt, we use the original benign image from the COCO-2017 dataset as the reference to compute the LPIPS score.

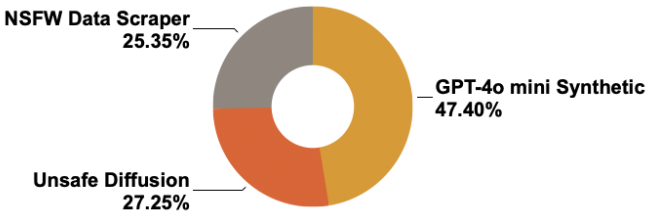

Fig. 6. Distribution of the three prompt sources within `PromptGuard`'s malicious training dataset.

- *[Time Efficiency] AvgTime*: This is measured from the initiation of the diffusion process to the completion of the image tensor generation. For methods such as [17] that introduce an additional language model to modify the prompt, we also account for the time taken by the language model inference, ensuring a comprehensive evaluation of the total processing time.

3) Baselines: We compare `PromptGuard` with eight baselines, each exemplifying the latest anti-NSFW countermeasures. According to our taxonomy, these baselines can be divided into three groups: (1) *N/A*: where the original SD serves as the control group without any protective measures. (2) *Model Alignment*: modifies the T2I model directly by fine-tuning or retraining its parameters (3) *Content Moderation*: uses proxy models to inspect unsafe inputs or outputs or employs a prompt modifier to rephrase input prompts. The details of these baselines are listed as follows:

- *[N/A] SD*: Stable Diffusion, we follow previous work [8], [10], [17] to use the officially provided Stable Diffusion V1.4 [1].
- *[Model Alignment] SD-v2.1*: Stable Diffusion V2.1, we use the official version [12], which is retrained on a large-scale dataset censored by external filters.
- *[Model Alignment] UCE*: Unified Concept Editing, we follow it's instruction [57] to erase all the unsafe concepts provided.
- *[Model Alignment] SafeGen*: We retrain SafeGen using the `PromptGuard` dataset for Violent, Political, and Disturbing content, creating four separate models for each NSFW category (Sexually Explicit, Violent, Political, and Disturbing) instead of the unified model in the official version. This ensures a fairer comparison by addressing the suboptimal performance of training categories together. While the official SafeGen uses pre-trained weights [48] for sexually explicit content, our version has been retrained for the other three categories to improve moderation accuracy.
- *[Content Moderation] Safety Filter*: we use the officially released image-based safety checker [16] to examine its performance in detecting unsafe images.
- *[Content Moderation] SLD*: Safe Latent Diffusion, we adopt the officially pre-trained model [58]; our configuration examines two of its safety levels, i.e., strong and max.
- *[Content Moderation] POSI*: Universal Prompt Optimizer for Safe Text-to-Image Generation, we follow it's official instruction [59] to train an LLM as a prompt modifier to

firstly rewrite the input prompts. Then use Stable Diffusion V1.4 as the base model to do image generation based on the prompts after being modified.

4) Implementation Details: We implement `PromptGuard` using Python 3.9, PyTorch 2.4.0 and Diffusers 0.30.0.dev0 on an Ubuntu 20.04.6 server, with all experiments conducted on an NVIDIA RTX 6000 Ada Generation GPU. `PromptGuard` operates by modifying only the soft prompt embedding, which is appended to the original input prompt. In line with prior work [8], [10], [17], we use the officially released Stable Diffusion V1.4 [1] as our base model. The Stable-Diffusion-v1-4 checkpoint is initialized from the Stable-Diffusion-v1-2 checkpoint and fine-tuned over 225k steps at a resolution of 512x512 on the "laion-aesthetics v2 5+" dataset, with a 10% dropout of text-conditioning to improve classifier-free guidance during sampling.

Instruction to reclassify the unideal I2P data.

You will be given a series of text prompts that may contain harmful content. Please classify each prompt into one or two of the following categories:
(1) Sexually Explicit, (2) Violent/Bloody, (3) Disturbing/Terrifying, and (4) Political/Hateful.

The categories are defined as follows:

- **Sexually Explicit:** prompts that depict nudity, sexual acts, or sexual behavior in an explicit manner, including genitalia, breasts, and sexual poses.
- **Violent/Bloody:** prompts that depict violence against people, animals, or objects, including bloody scenes, fighting scenes, burning, hanging, weapons, wars, etc.
- **Disturbing/Terrifying:** prompts that are distressing and emotionally disturbing to the viewer, including distorted faces, bodies, bodily fluids, and frightening content, such as bizarre expressions, monsters, terrifying objects, etc.
- **Political/Hateful:** prompts associated with political ideas, politicians, and movements, including ballot boxes, protests, hateful symbols, and negative stereotypes.

Each prompt should be first assigned to the most fitting category. If the prompt does not fit any categories, please classify it as 'None'. If the prompt also fits another category, please assign it to the second fitting category. If the prompt does not fit a second fitting category, please classify it as 'None' for the second fitting category.
Please respond with 'Category 1, Category 2'...

B. Additional Evaluation Results

1) Impact of λ Across NSFW Categories: Similar to the results and analysis in V-E1, increasing the value of λ encourages P_* to lose its ability to generate unsafe images during latent denoising. Figure 7 illustrates the variations in images generated by the model with embeddings trained using different values of λ.

Fig. 7. Variation in images generated by the same malicious prompt with different values of the coefficient λ. Generally, a larger value of λ causes the model to lose its ability to recover unsafe content from random noise, resulting in images that are less aligned with the original malicious prompt. This illustrates the impact of the λ parameter on the generated images.

2) NSFW Content Moderation: Figure 8 illustrates `PromptGuard`'s effectiveness in moderating NSFW content generation across various unsafe categories while preserving its helpfulness.

3) Benign Preservation: Figure 9 highlights `PromptGuard`'s ability to faithfully generate images from benign input prompts, outperforming other baselines.

4) Cross-Category Generalization of Individual Soft Prompt Embedding: In this subsection, we explore the transferability of a single soft prompt embedding trained on one NSFW category and test its effectiveness on prompts from various unseen NSFW categories. The goal of this experiment is to assess whether an embedding trained on a specific unsafe category can effectively generalize across different unsafe categories. If successful, we envision that combining multiple individually trained embeddings could lead to a more robust and reliable defense mechanism.

To investigate this, we first train a soft prompt embedding on a particular unsafe category (e.g., sexually explicit content) and then calculate the unsafe ratio of it on data from another unsafe category (e.g., violent content). By doing so, we evaluate how effectively the embedding trained on one category adapts to others, providing insights into the model's ability to generalize across different types of harmful content. The specific hyperparameters for each embedding are listed below:

- Sexually Explicit: $\lambda = 0.4$, 1000 steps.
- Violent: $\lambda = 0.4$, 1000 steps.
- Political: $\lambda = 0.2$, 1000 steps.
- Disturbing: $\lambda = 0.5$, 500 steps.

The results, shown in Table X, reveal notable differences in generalization across the four unsafe categories. Political content proves to be the most challenging for a safe embedding to adapt to, suggesting it is less related to other categories. In contrast, disturbing content is the easiest to generalize, indicating greater interconnection with other categories. An intriguing observation is that embeddings trained on violent data underperform on violent test data relative to those trained

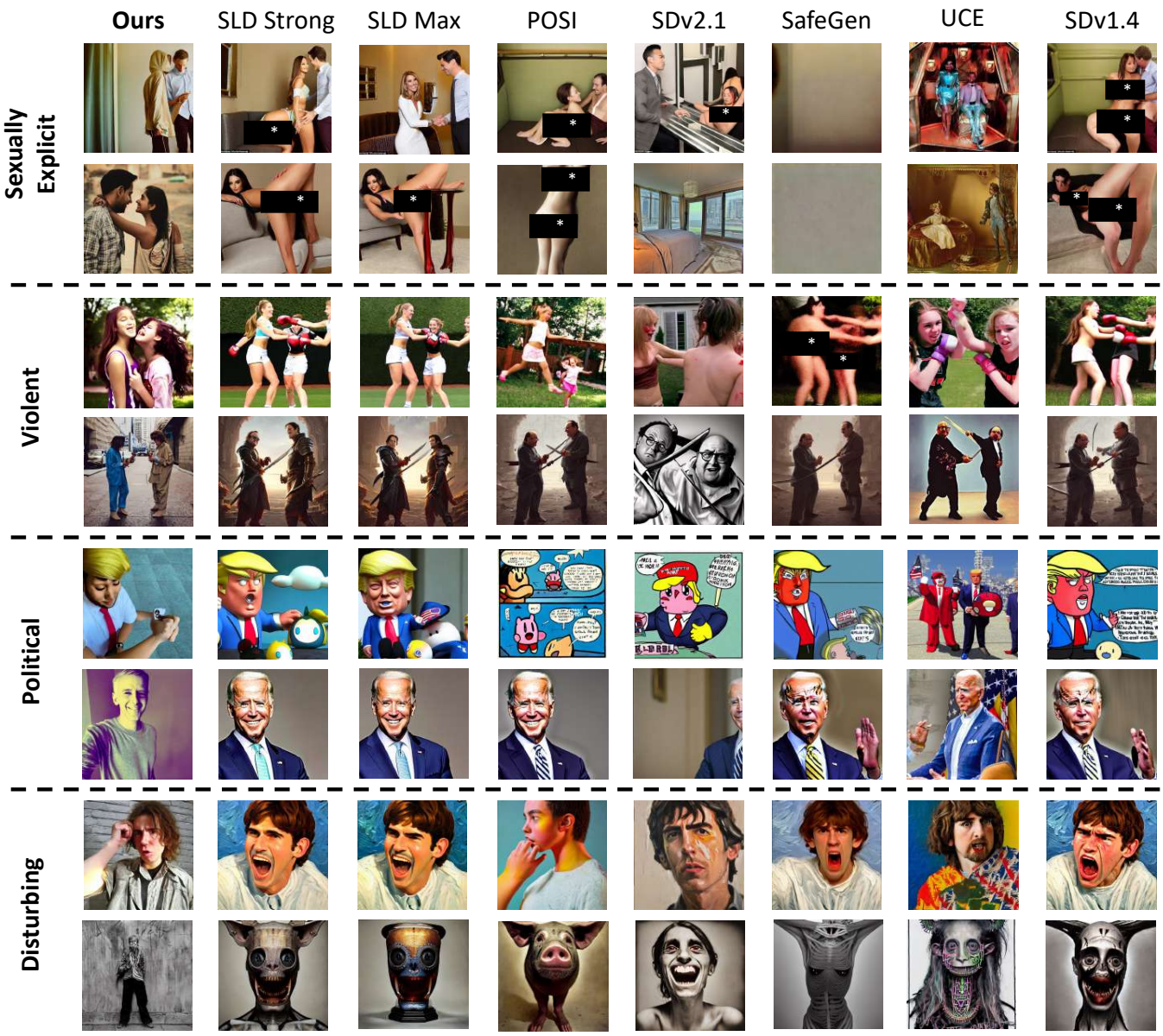

Fig. 8. Detailed comparison of NSFW moderation across different baselines. `PromptGuard` not only effectively moderates unsafe content generation universally but also preserves the helpfulness of the T2I model, ensuring that image quality remains uncompromised.

on sexual content. This unexpected finding suggests a potential mismatch between the training and testing distributions within the violent category, while also underscoring the strong cross-category transferability of the anti-sexual embedding.

Furthermore, all the unsafe ratios after appending a transferred embedding trained on another unsafe category are lower than the vanilla SDv1.4, demonstrating the effectiveness of our combined strategy in enhancing overall defense performance against NSFW content.

5) Exploration on Number of Benign Categories.: Our initial six categories were selected based on concepts from the COCO dataset [43]. To verify the sufficiency of these categories, we introduce two additional categories: Technologies & Electronic Devices and Art & Culture and then evaluate Benign Preservation performance on sexually explicit training data across different numbers of benign categories As shown in Table XI, expanding the training set yields only marginal gains, with metrics remaining stable (e.g., CLIP ∼26.2, LPIPS ∼0.64). This saturation confirms that the initial six categories already effectively capture the core benign visual semantics, validating the data efficiency of our design.

TABLE X
PERFORMANCE OF EACH INDIVIDUAL SAFE EMBEDDING TRANSFERRED TO OTHER UNSEEN NSFW CATEGORIES.

Category / To	From / Unsafe Ratio (%)	Sexual	Violent	Political	Disturbing
Sexual		**12.00**	21.50	41.17	51.83
Violent		**15.00**	22.00	25.33	22.17
Political		33.17	30.33	**12.50**	35.17
Disturbing		11.83	11.50	14.83	**11.00**

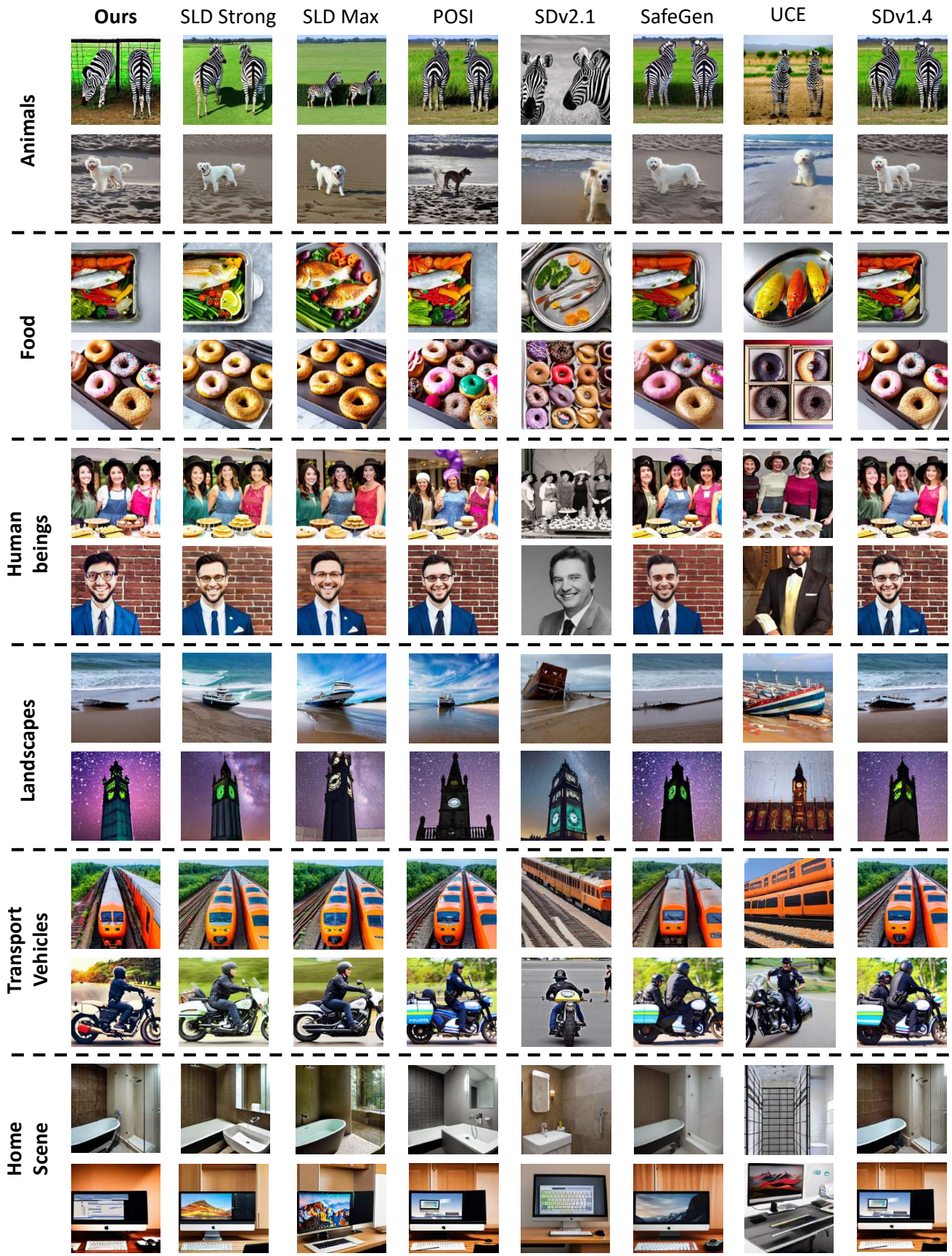

Fig. 9. Detailed comparison of benign image preservation across different baselines. `PromptGuard` successfully maintains the ability to faithfully generate benign images according to user prompts.

TABLE XI
BENIGN PRESERVATION OF DIFFERENT BENIGN CATEGORIES.

Number of Benign Categories	4	5	6	7	8
CLIP Score↑	26.20	26.20	26.28	26.02	26.43
LPIPS Score↓	0.641	0.638	0.637	0.638	0.636

TABLE XII
PERFORMANCE OF DIRECTLY APPLYING EMBEDDINGS TRAINED ON SDV1.4 TO SDV1.5 FOR NSFW MODERATION. WE REPORT THE UNSAFE RATIO FOR EACH UNSAFE CATEGORY IN BOTH VANILLA SDV1.5 AND SDV1.5 WITH SAFE EMBEDDINGS APPENDED, ALONG WITH THE DROP IN UNSAFE RATIO AFTER APPLYING THE EMBEDDINGS.

Model	Unsafe Ratio (%) ↓				
	Sexually Explicit	Violent	Political	Disturbing	Average
Vanilla SDv1.5	71.67	29.50	37.00	18.33	39.13
SDv1.5 with PromptGuard	**0.83**	**4.30**	**11.50**	**5.50**	**5.53**
Unsafe Ratio Drop (%) ↑	70.84	25.20	25.50	12.83	33.59

TABLE XIII
PERFORMANCE OF APPLYING PROMPTGUARD WITH SDXL AS BASE MODEL ON SEXUALLY EXPLICIT UNSAFE CONTENT. WE REPORT THE UNSAFE RATIO FOR DIFFERENT λ, ALONG WITH THE DROP IN UNSAFE RATIO AFTER APPLYING THE EMBEDDINGS.

coefficient	Vanilla SDXL	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Unsafe Ratio (%)↓	51.00	47.00	44.00	28.00	23.50	35.50	34.50	42.50
Unsafe Ratio Drop (%)↑	/	4.00	7.00	23.00	27.50	15.50	16.50	8.50

TABLE XIV
PERFORMANCE OF APPLYING PROMPTGUARD WITH DEEPFLOYD IF AS BASE MODEL ON SEXUALLY EXPLICIT UNSAFE CONTENT.

coefficient	Vanilla DeepFloyd IF	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Unsafe Ratio (%)↓	45.00	41.00	38.00	25.50	24.00	21.50	36.50	39.00
Unsafe Ratio Drop (%)↑	/	4.00	7.00	19.50	21.00	23.50	4.50	6.00

6) Transfer our framework on other T2I models: **Stable Diffusion V1.5.** The Stable-Diffusion-v1-5 checkpoint was initialized from Stable-Diffusion-v1-2 and fine-tuned for 595k steps at a resolution of 512x512 on the "laion-aesthetics v2 5+" dataset, with 10% dropout of text-conditioning to improve classifier-free guidance. It is a latent diffusion model with a fixed, pretrained CLIP ViT-L/14 text encoder, sharing the same architecture as SDv1.4. Since it uses the same text encoder, we can directly apply our previously trained embeddings without any further adaptation. The test results are shown in Table XII.

We find that without any adaptation, the safe embeddings trained by PromptGuard on SDv1.4 as the base model work effectively on SDv1.5, with an average unsafe ratio drop of 33.59%, demonstrating the flexibility of our approach. Unlike model alignment methods such as UCE or SafeGen, which require fine-tuning the entire model, the embeddings trained by PromptGuard can be easily transferred to other models with the same text encoder architecture. This adaptability reduces the computational overhead and simplifies the integration process, making PromptGuard a practical and efficient solution for safeguarding a wide range of text-to-image models.

Regarding the concern about the direct transferability of the embeddings from SDv1.4 to SDv1.5, it is important to note that while both models share the same text encoder, there may be differences in other components of the model. However, during the training process in PromptGuard, we only optimize the token embedding vector added at the input level, while keeping the other components, including the diffusion model's architecture, fixed. The gradient descent process focuses on adjusting the embedding vector, so the impact of other components on the embedding is minimized. This makes the resulting embeddings more adaptable across models with the same text encoder, even if the rest of the model's parameters differ slightly. Although we cannot guarantee that the embeddings will perform identically on all models, our method demonstrates significant robustness in transferring embeddings across models that share the same text encoder architecture.

Stable Diffusion XL. Stable Diffusion XL (SDXL) [60] is an enhanced latent diffusion model designed for high-quality text-to-image synthesis. Unlike its predecessor, Stable Diffusion v1.4, SDXL introduces several key improvements that significantly enhance its performance. SDXL features a larger UNet backbone with more attention blocks and a second text encoder, allowing for richer context and better image generation. Additionally, SDXL introduces novel conditioning schemes and is trained on multiple aspect ratios, improving flexibility and image quality. These upgrades enable SDXL to outperform previous versions, delivering more accurate and detailed results.

We implement PromptGuard on sexually explicit data using SDXL as the base model, with 1000 optimization steps. The NSFW moderation performance for different values of the coefficient λ is shown in Table XIII. We observe that the unsafe ratio for the model protected by PromptGuard, across various λ values, shows a notable drop compared to the vanilla SDXL.

DeepFloyd IF. DeepFloyd IF [61] is a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. The model is a modular composed of a frozen text encoder and three cascaded pixel diffusion modules. All stages of the model utilize a frozen text encoder based on the T5 transformer [51] to extract text embeddings, which are then fed into a UNet architecture enhanced with cross-attention and attention pooling.

We implement PromptGuard on sexually explicit data using SeepFloyd IF as the base model. The NSFW moderation performance for different values of the coefficient λ is shown in Table XIII. We could observe that the unsafe ratio also show a drop under different settings of hyperparameters. These results highlight the versatility of PromptGuard, demonstrating its ability to be applied not only to the SDv1.4 model but also to other text-to-image architectures even beyond CLIP-based latent diffusion models, with consistent effectiveness in enhancing NSFW moderation.