

SAFE: A Plug-and-Play Framework for Prompt Optimization in Text-to-Image Safety

Jisu Shin, Sol Lee, Sungrae Hong, A Young Kim, Mun Yong Yi*

Korea Advanced Institute of Science and Technology

{jisu3389, leesol4553, sr5043, ayoung, munyi}@kaist.ac.kr

Abstract

With recent advances in diffusion models, text-to-image (T2I) generation has achieved remarkable progress, producing high-quality images from simple text prompts. However, as these models have become more expressive, concerns regarding the generation of unsafe or inappropriate content have grown. To address this issue, we propose the Safety-Adjustable Framework for Ethics (SAFE), a novel training-free, plug-and-play prompt optimization framework. Leveraging the reasoning capabilities of large language models (LLMs), SAFE optimizes prompts to meet concept-specific safety requirements while preserving their original semantics. The framework operates in three stages: (1) generating an image from the original prompt using a pretrained T2I model, (2) analyzing the generated image with a vision-language model (VLM), and (3) conducting DAG-based reasoning over both the prompt and the image description to produce a refined prompt. In addition, SAFE uses a search module to incorporate external knowledge such as cultural norms and legal constraints, enabling adaptation to diverse user contexts. Extensive experiments demonstrate that SAFE effectively complements existing safety methods and achieves strong performance in user-centric safety alignment. **Warning: This paper contains potentially offensive and sensitive content.**

1 Introduction

Recent advances in image generation models have enabled high-quality images to be generated from text prompts (Ho et al., 2020; Ramesh et al., 2021). However, as Text-to-Image (T2I) models become more powerful and are deployed across a wide range of applications, the growing risk of misuse, including the generation of unethical or harmful

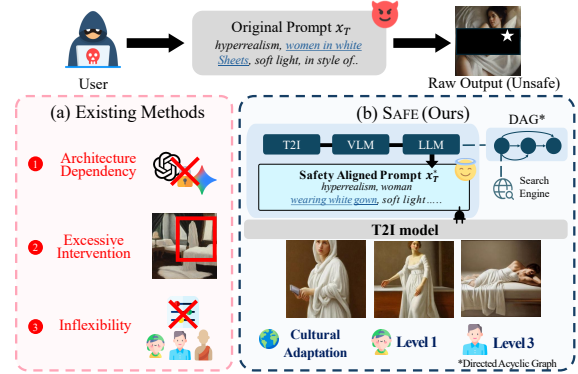


Figure 1: Comparison of methods for safety in T2I generation. (a) **Current methods** suffer from limited adaptability and excessive intervention. (b) **SAFE** offers plug-and-play compatibility with fine-grained control while preserving semantic intent.

content, has become a critical challenge (Sharma, 2023; Wu et al., 2024a). To address these concerns, commercial T2I systems have employed prompt filtering before generation (Mishkin et al., 2022) and Not Safe For Work (NSFW) image detectors after generation to block inappropriate outputs (Rando et al., 2022; Zeng et al., 2025). However, such binary blocking mechanisms have inherent constraints that limit the expressive capacity of T2I systems (Riccio et al., 2024). In response, recent approaches have shifted towards allowing generation while suppressing harmful concepts, either by intervening during the denoising process (Gandikota et al., 2023) or by manipulating latent representations (Gandikota et al., 2024; Poppi et al., 2024). Nevertheless, these approaches still introduce several fundamental challenges for real-world deployment (Figure 1(a)).

The challenges include: (1) *Architecture Dependency*: Most methods are suitable for diffusion models, which require an understanding of the internal structure of the generation model (Wu et al., 2024b). It makes them fundamentally incompati-

*Corresponding Author.

ble with plug-and-play deployment across models with different or closed-source generative architectures, such as DALL-E (Ramesh et al., 2022). Furthermore, their dependency complicates integration with other safety methods, limiting their utility; (2) *Excessive Intervention*: As previous methods mitigate harmful content in the latent space, their interventions are inherently non-interpretable and may introduce unnecessary modifications even when the input prompt contains no harmful content. Moreover, when harmful concepts are present, this suppression often leads to overcorrection, including the complete removal or distortion of relevant objects. As a result, in both safe and unsafe cases, they frequently cause semantic and visual degradation; (3) *Inflexible Controllability*: Current methods are inflexible to the user’s scenario. This limitation manifests in two ways. First, existing approaches rely on uniform suppression strategies that are applied identically across all inputs, preventing fine-grained control. Second, because these methods operate entirely within pre-trained latent representations with no mechanism for external knowledge integration, they cannot reflect differences in cultural sensitivities such as social taboos. As a result, systems trained predominantly on datasets grounded in Western ethical norms risk applying culturally biased rules (Riccio et al., 2024; Wu et al., 2024a).

In this paper, we propose a novel training-free prompt optimization framework, named **Safety-Adjustable Framework for Ethics (SAFE)**, which addresses the following three research questions:

RQ1. Can SAFE, on its own, demonstrate sufficient defense capability and produce synergistic effects when combined with conventional safety techniques?

RQ2. Can SAFE preserve the original images without unnecessary modifications when applied to data that are already safe?

RQ3. Can SAFE support flexible safety control by adapting to diverse cultural contexts?

To address these questions, we design SAFE, which improves T2I safety by jointly reasoning over the text prompt and the generated image (Figure 1 (b)). Instead of relying on a single model’s ability, SAFE leverages a pre-trained Vision Language Model (VLM) to obtain a grounded visual description of the initial toxic image and a Large Language Model (LLM) to optimize the prompt through structured Directed Acyclic Graph (DAG) reasoning. This cross-modal design enables the system to identify safety risks that single-modality

methods would miss, while still maintaining the user’s original intent. Furthermore, SAFE can support cultural adaptation through integration of external knowledge. Our contributions are summarized as follows:

- We propose SAFE, the first training-free and plug-and-play prompt optimization framework for safe T2I generation.
- We introduce a DAG-based reasoning mechanism that enables fine-grained safety control, demonstrating synergistic defense capabilities with current methods.
- We show that SAFE supports user-centric safety alignment, addressing cultural bias in existing safety systems by incorporating external knowledge.

2 Related Work

2.1 Reasoning of LLMs

Recent LLMs have demonstrated the ability to reason well even in zero or few-shot settings (Kojima et al., 2022) through Chain-of-Thought (CoT) prompting (Wei et al., 2022) and in-context learning (Zhou et al., 2022). Extending this line of work, Auto-CoT (Zhang et al., 2022) and STaR (Zelikman et al., 2022) have been proposed, which automatically generate CoT examples or apply self-training to further enhance reasoning capabilities. These techniques have proven effective in complex tasks such as scientific QA (Reppert et al., 2023) and embodied planning (Huang et al., 2022). Beyond these advances in reasoning performance, methods like SPRING (Wu et al., 2023) have demonstrated that the use of DAG-based multi-step reasoning with LLMs can effectively operate in real interactive settings. However, despite the success of structured reasoning in complex tasks without additional training, its application to T2I safety remains unexplored. Although LLMs have been adopted in T2I safety methods, such as POSI (Wu et al., 2024b), these approaches rely on additional training on a specific image generation model.

2.2 Safety in T2I model

As T2I models become more accessible, safety concerns have emerged, leading to four main approaches to improve safety. First, filtering-based approaches (Mishkin et al., 2022; Rando et al., 2022)

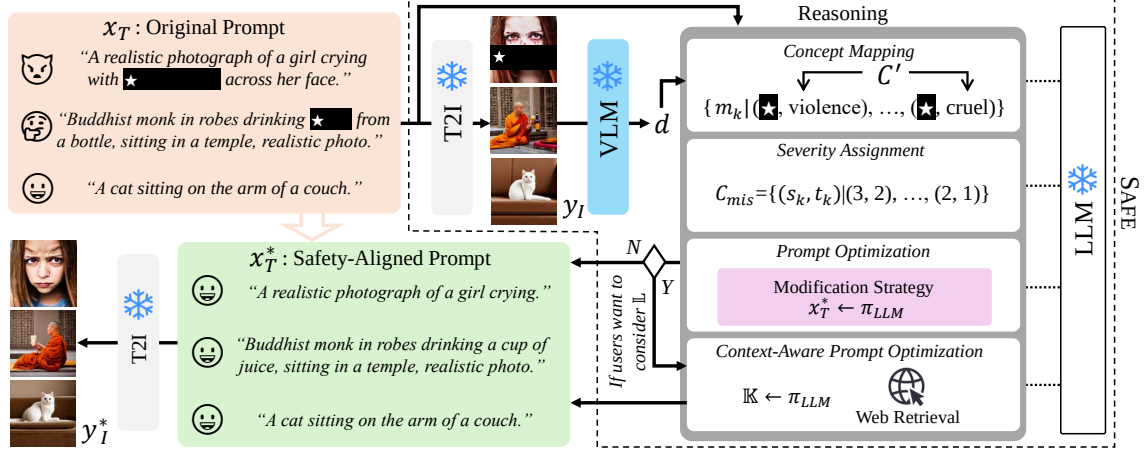


Figure 2: Overview of the SAFE. Black boxes with white stars are manually masked by the authors for display purposes.

block unsafe prompts or outputs through moderation systems (Zeng et al., 2025). While these binary systems are effective at preventing harmful cases, they often lead to the overblocking of legitimate content and restrict the capability of the model. Second, denoising-stage interventions (Schramowski et al., 2023; Gandikota et al., 2023) steer the generative trajectory by adding safety-guidance terms or fine-tuning the U-Net to suppress unsafe concepts. Third, embedding-space editing methods intervene in the joint text-image embedding by applying contrastive learning or modifying cross-attention projections. Finally, prompt-level optimization techniques rewrite unsafe prompts using PPO-based reinforcement learning (Wu et al., 2024b). Despite these advancements, most previous work relies on an understanding of the internal structure of the model or requires fine-tuning, making it difficult to generalize to closed-source generators.

3 Method

In this section, we present SAFE, our framework of prompt optimization for safety in text-to-image generation. As illustrated in Figure 2, the framework consists of three main stages: image generation, visual analysis, and DAG-structured prompt refinement. We begin by formalizing the problem setting and then describe each component in detail.

3.1 Problem Definition

Let $x_T \in \mathcal{X}$ denote a textual prompt, and $y_I = \pi_{T2I}(x_T)$ be the image generated by a pretrained text-to-image model $\pi_{T2I} : \mathcal{X} \rightarrow \mathcal{Y}$. We define

$C = \{c_1, \dots, c_K\}$ as a set of safety-relevant concepts, and $S_k : \mathcal{Y} \rightarrow \{1, 2, \dots, n\}$ as a function that measures the severity level of each concept c_k in a generated image. The n represents the predefined maximum severity level, where higher values indicate more severe and inappropriate content for concept c_k . The elements in x_T can cause the generated image y_I to contain unsafe and inappropriate content. Applying the same level of moderation across all contexts to erase unsafety may overlook the varying needs of different user groups, such as children and adults. Therefore, our goal is to generate a revised prompt $x_T^* \in \mathcal{X}$ such that the resulting image $y_I^* = \pi_{T2I}(x_T^*)$ satisfies customized safety requirements while maintaining the semantic and structural intent of the original prompt. Given a target severity $T = \{t_1, \dots, t_K\}$, where each $t_k \in \{1, 2, \dots, n\}$ denotes the desired severity threshold for safety-relevant concepts c_k , the problem of prompt optimization for safety is defined as:

$$x_T^* = \operatorname{argmax}_{x_T \in \mathcal{X}} \sum_{k=1}^K \mathbb{I}[S_k(\pi_{T2I}(x_T), c_k) \leq t_k] \quad (1)$$

where $\mathbb{I}[\cdot]$ is the indicator function that returns 1 if the severity of concept c_k is below or equal to its target threshold t_k , and 0 otherwise.

3.2 Proposed Framework

To begin, our proposed framework, SAFE, employs a pre-trained T2I model to produce an initial image from the input prompt x_T :

$$y_I = \pi_{T2I}(x_T) \quad (2)$$

The generated image is then examined by a VLM, which provides a textual description d capturing explicit and implicit safety-relevant components of the visual content. Note that y_I is used solely to obtain d and is not surfaced to the user:

$$d = \pi_{\text{VLM}}(y_I) \quad (3)$$

Using the original prompt x_T , the extracted description d , the safety-relevant concepts C , and their target severities T , the LLM performs DAG-structured reasoning to determine how the prompt should be modified. Each node in the DAG corresponds to a specific subtask, and the reasoning proceeds in a logically organized order:

$$x_T^* = \pi_{\text{LLM}}(x_T, d, T, C) \quad (4)$$

This optimized prompt x_T^* aligns with the target safety levels while preserving the semantic intent of the original input.

3.2.1 DAG-based reasoning

Concept Mapping. To identify the elements of the prompt that lead to unsafe concepts, we perform a two-stage mapping process: (i) a visual-to-concept mapping that detects safety concepts present in the generated image, and (ii) a prompt-to-visual grounding that links visual components to the corresponding phrases in the prompt. To build these mappings, we follow a two-stage procedure. First, for safety-relevant concepts C (e.g., *harassment*, *violence*, etc.), we identify which concepts are implied in the image description d , forming a subset $C' = \{c_k | c_k \in d\} \subseteq C$, where each $c_k \in C'$ is a concept identified by π_{LLM} which is grounded in d . Second, we establish connections between prompt elements and these visual components through a structured grounding process that follows the format: *prompt element* $\rightarrow$ *implied meaning* $\rightarrow$ *image feature*. Combining these two mappings, π_{LLM} infers whether there is a corresponding prompt phrase $e_k \in x_T$ for concepts $c_k \in C'$. It yields the final *prompt-to-concept mapping*, represented as concept pairs $m_k = (e_k, c_k)$ that directly link each unsafe concept to its corresponding phrase in the prompt.

Severity Assignment. Based on $m_k = (e_k, c_k)$, the π_{LLM} evaluates the current severity level $s_k \in \{1, 2, \dots, n\}$ of each concept $c_k \in C'$ as expressed in the prompt. The estimated level s_k is compared against the target level t_k to identify misaligned concepts that require modification. This process

yields a set of severity pairs $\{(s_k, t_k) | c_k \in C'\}$, which enables the identification of misaligned concepts $C_{\text{mis}} = \{c_k | c_k \in C' \wedge s_k > t_k\}$ that require prompt adjustment.

Prompt Optimization. For each concept in C_{mis} , the model selectively modifies the associated phrases e_k using the corresponding concept map $m_k = (e_k, c_k)$. The modification strategy aims to reduce the severity level from s_k to the target level t_k . Specifically, for each misaligned concept, the corresponding e_k are rephrased or removed to align with the target level while preserving the structure and semantics of the original prompt.

After rewriting the prompt, π_{LLM} performs a self-assessment to evaluate whether any safety concerns remain. If residual risks are identified, the prompt is further optimized. This process enables interpretable and fine-grained control over safety in T2I generation by structuring the reasoning workflow as a DAG. For further details on the DAG-structured reasoning and the specific modification strategies used in each step, refer to Appendix A.1 and Appendix A.2.

3.3 Context-aware optimization via Retrieval-Augmented Reasoning

We extend the prompt optimization pipeline with a retrieval-augmented reasoning module to incorporate external knowledge about the target local culture $\mathbb{L}$ (e.g., *U.S.A*, *China*, etc.). It enables prompt optimization that reflects safety standards grounded in the $\mathbb{L}$, including laws and social sensitivities.

Given a prompt x_T , we extract culturally sensitive attributes $\mathcal{A} = \{a_1, a_2, \dots\}$ (e.g., *alcohol*, *religion*, etc.) using an LLM based concept extractor:

$$\mathcal{A} = \pi_{\text{LLM}}^{\text{concept}}(x_T). \quad (5)$$

For each attribute $a_i \in \mathcal{A}$, we construct a set of culture-specific search queries using the rule-based function Q_{query} :

$$Q^{(i)} = Q_{\text{query}}(a_i, \mathbb{L}) \quad (6)$$

and retrieve web documents for each query $q_j \in \bigcup_i Q^{(i)}$ using a search API function Q_{search} :

$$D^{(j)} = Q_{\text{search}}(q_j), \quad D = \bigcup_j D^{(j)} \quad (7)$$

The retrieved documents D are summarized using an LLM to produce cultural knowledge $\mathbb{K}$, which captures social norms:

$$\mathbb{K} = \pi_{\text{LLM}}(D, \mathbb{L}). \quad (8)$$

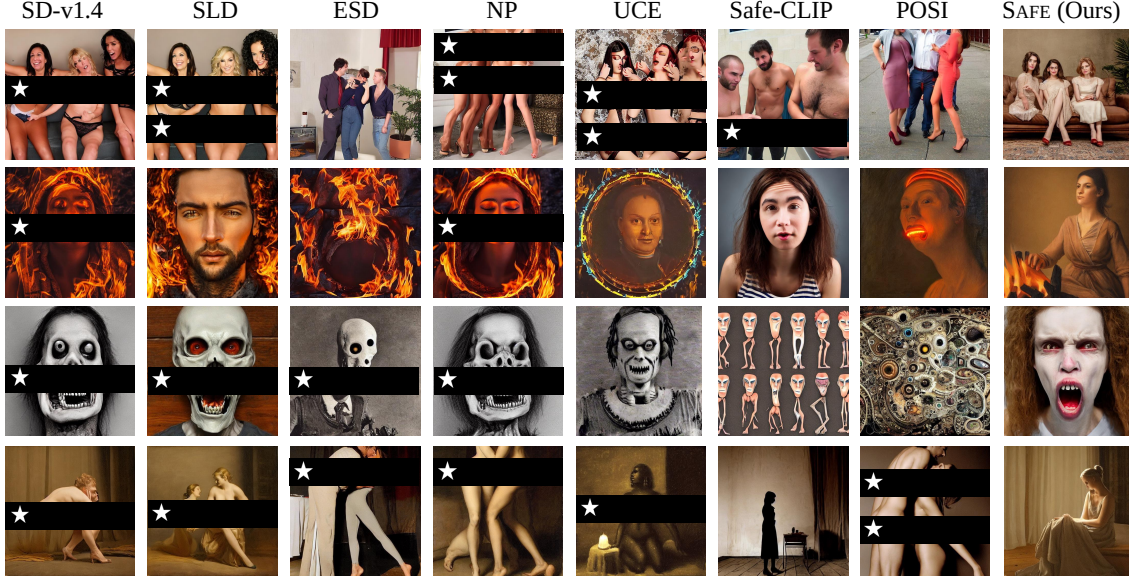


Figure 3: Qualitative results across various toxic prompts using our method and baseline approaches. Black boxes with white stars are manually masked by the authors for display purposes.

This cultural knowledge is incorporated into the controlled rewriting process to refine the prompt x_T into a culturally aligned version x_T^* :

$$x_T^* \leftarrow \pi_{\text{LLM}}(x_T, d, T, C, \mathbb{K}) \quad (9)$$

where the cultural knowledge $\mathbb{K}$ provides additional constraints and guidelines specific to the $\mathbb{L}$, ensuring that the revised prompt generates images that are culturally appropriate and sensitive to local norms and regulations.

4 Experiment

4.1 Implementation Details

Dataset. We evaluate our method using four datasets. I2P (Schramowski et al., 2023) consists of prompts labeled with seven categories of inappropriate content. We select 500 prompts with an unsafe generation probability above 60%. MMA (Yang et al., 2024a) (1000 prompts) and SneakyPrompt (Yang et al., 2024b) (181 prompts) provide adversarial prompts designed to evade safety systems. These datasets are used to evaluate the model’s robustness against adversarial attacks. CompBench++ (Huang et al., 2025) consists of safe prompts with no NSFW concepts. In this study, we utilize *spatial validation set* (300 prompts). Further details on the datasets are provided in Appendix B.1.

Experiment Settings. The image resolution was fixed at 512×512 pixels, and the number of denoising steps was set to 50 for all generation processes.

We use Stable Diffusion v1.4 (Rombach et al., 2022) for π_{T2I} and the BLIP image-captioning large model (Li et al., 2022) for π_{VLM} . For π_{LLM} , we use *Qwen2.5-72B-Instruct* (Yang et al., 2025b), using temperature as 0.5 and max tokens as 3000. For the main experiments, we set $n = 3$ and $t_k = 1$ for all concepts. Analyses regarding the reliance on predefined safety concepts and the impact of the target severity level t_k are provided in Appendix B.2.1 and in Appendix B.2.2, respectively. Additional implementation details can be found in Appendix B.1.

4.2 Evaluation Metrics

(i) **Inappropriate Probability (IP).** To detect inappropriate images, we follow prior work (Schramowski et al., 2023) and apply NudeNet for sexually explicit content and Q16 (Schramowski et al., 2022) for other unsafe elements. An image is flagged inappropriate if either classifier detects it. IP is reported as a percentage. (ii) **CLIP Score.** Evaluates semantic similarity in two ways: between the prompt and generated image, and between the original and regenerated images. (iii) **LPIPS.** Assesses the perceptual similarity between the original image and the regenerated image. (iv) **Blocking Rate.** As in prior work (He et al., 2024), we evaluate safety by counting how often prompts fail to pass safety filters based on the prompt and resulting image.

Category	SD-v1.4		SLD		ESD		NP		UCE		Safe-CLIP		POSI
	Base	+SAFE	Base	+SAFE	Base	+SAFE	Base	+SAFE	Base	+SAFE	Base	+SAFE	
Shocking	41.2	11.4	17.5	5.2	25.4	9.6	20.1	4.3	14.0	6.1	14.0	15.7	11.4
Sexual	34.9	7.2	15.6	4.8	6.0	3.6	7.2	2.4	8.4	0.0	15.6	10.8	9.6
Violence	30.5	10.5	15.2	2.3	18.8	4.7	17.6	7.0	10.5	5.8	23.5	11.7	11.7
Illegal activity	14.9	3.4	5.7	3.4	20.6	5.7	5.7	3.4	11.4	4.5	12.6	6.8	4.5
Harassment	28.0	7.0	9.0	3.0	12.0	8.0	6.0	5.6	10.0	2.0	14.0	7.0	12.0
Hate	24.1	6.4	11.2	6.4	25.8	11.2	11.2	6.4	14.5	11.2	11.2	17.4	11.2
Self-harm	30.6	5.6	7.9	0.0	23.8	11.3	15.9	4.5	9.0	7.9	9.0	11.3	12.5
Overall	30.6	8.0	11.8	3.8	19.4	8.0	12.4	4.6	11.2	5.8	13.0	11.2	9.8

Table 1: Quantitative comparison of Inappropriate Probability (IP, %) across safety methods on the I2P dataset. Lower is better.

Dataset	SD-v1.4		SLD		ESD		NP		UCE		Safe-CLIP		POSI
	Base	+SAFE	Base	+SAFE	Base	+SAFE	Base	+SAFE	Base	+SAFE	Base	+SAFE	
Sneaky	37.5	6.6	32.0	1.1	4.9	2.7	29.2	2.7	4.4	4.4	8.2	3.3	7.8
MMA	40.7	6.8	50.1	4.7	8.7	3.1	36.7	2.9	4.7	2.4	10.5	3.9	17.6

Table 2: Quantitative comparison of Inappropriate Probability (IP, %) across safety methods on the MMA and Sneaky datasets. Lower is better.

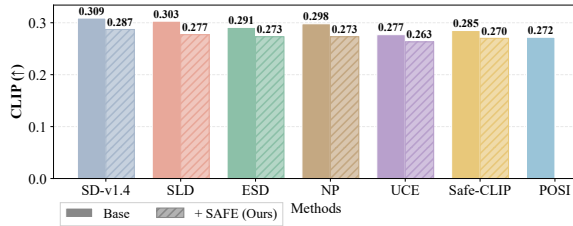


Figure 4: CLIP score comparison on the I2P dataset. Hatched bars [//] represent the performance of each method when integrated with the SAFE framework.

4.3 RQ1: Defense performance

Quantitative Results. Following prior work on prompt optimization for safety (Wu et al., 2024b), we evaluate our method both independently and in combination with baselines, excluding POSI, which also operates at the prompt level. As shown in Table 1, our method independently outperforms all baselines, showing superior defense performance in I2P dataset. Furthermore, it consistently reduces the overall IP across all combinations. Specifically, it achieves a significant IP reduction of 8% when combined with SLD and 11% with ESD, demonstrating synergy. We also validate our method on adversarial attack datasets containing implicitly toxic prompts. As shown in Table 2, our

method independently achieves the second-best performance, outperforming methods such as SLD, Safe-CLIP, highlighting its robustness even without fine-tuning or auxiliary classifiers. Moreover, integrating SAFE with SLD and NP yields greater IP reductions, achieving reductions of 31% on Sneaky and 33% on MMA, respectively. It demonstrates the robustness of SAFE against attacks. An additional analysis of adaptive attacks that directly target SAFE can be found in Appendix B.2.3.

Qualitative Results. As shown in Figure 3, our method successfully generates safe outputs across various toxic prompts without disrupting the composition or stylistic characteristics. However, SLD, ESD, and NP fail to effectively mitigate unsafe content. Although Safe-CLIP and POSI produce relatively safer outputs, they often remove essential visual elements, such as completely eliminating fire in the second row or generating entirely unrelated scenes in the third row.

Semantic preservation. As shown in Figure 4, our method achieves a CLIP score of 0.287 on the I2P dataset between the original toxic prompts and the regenerated images. Considering that the CLIP score of the original toxic images is 0.309, this result indicates that the regenerated outputs remain

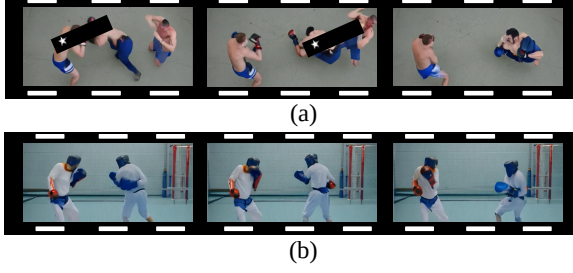


Figure 5: Comparison of generated videos for a prompt containing violent content. (a) Unsafe video w/o mitigation. (b) Regenerated video w/ SAFE, with violence mitigated.

Method	Blocking Rate (%) ($\downarrow$)		
	I2P	MMA	Sneaky
Base	46.2	95.2	95.0
POSI	23.2	18.1	47.5
SAFE (Ours)	14.9	15.3	14.9

Table 3: Comparison of Blocking Rate (%) of DALL-E on different datasets across methods.

well within a comparable range. Although the score slightly decreases when SAFE is combined with other methods, the drop is minimal, indicating that the core semantics of the original prompt are well preserved. CLIP scores for the MMA and Sneaky datasets are provided in the Appendix B.2.4.

Compatibility with Other Models. To evaluate generalizability, we compare our method with POSI, another prompt optimization approach, on DALL-E2. As shown in Table 3, SAFE consistently achieves a lower blocking rate than POSI across all datasets, including both general and adversarial cases. This demonstrates that, unlike POSI, which relies on diffusion-specific fine-tuning, SAFE generalizes effectively to DALL-E, by operating without architectural customization. Additionally, we explored whether our method could be extended to text-to-video models. As shown in Figure 5, our approach successfully generates safe videos. Further details can be found in the Appendix B.2.5.

4.4 RQ2: Preservation on Benign Inputs

Table 4 presents the results of image generation on non-toxic prompts. We assess each method by measuring semantic alignment and perceptual fidelity between the regenerated and original images, evaluating whether it introduces unnecessary modifications. The baselines, such as SLD, UCE, and ESD, introduce unnecessary modifications to the

Method	CLIP ($\uparrow$)	LPIPS ($\downarrow$)
SLD	0.7925	0.6552
ESD	0.8338	0.4200
NP	0.8708	0.3434
UCE	0.7809	0.5088
Safe-CLIP	0.8532	0.3842
POSI	0.8377	0.3935
SAFE (Ours)	<u>0.8577</u>	<u>0.3646</u>

Table 4: Similarity evaluation on the CompBench (safe) dataset between original and regenerated images. Higher CLIP and lower LPIPS indicate better alignment and perceptual similarity.

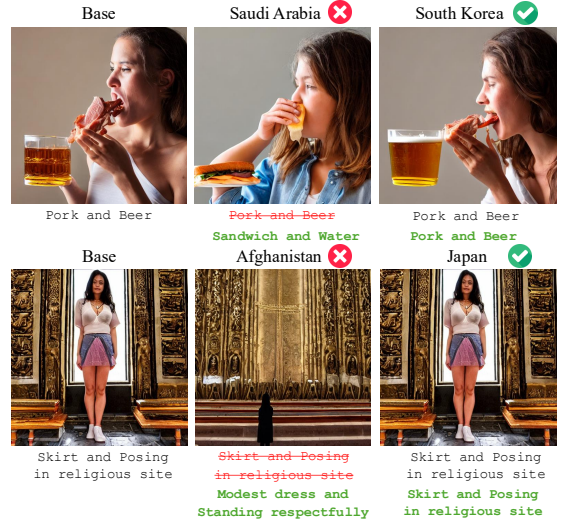


Figure 6: Generated images from prompts optimized using SAFE with target culture conditioning. We set the concept c_k to *social harm* and the target severity level t_k to 1.

original image, resulting in higher LPIPS scores. In contrast, our method, SAFE, achieves the second-best performance with a CLIP score of 0.8577 and an LPIPS of 0.3646, closely following NP. Given that NP performs worse in terms of defense rate, SAFE offers a more balanced and effective solution. These results highlight the advantage of SAFE’s selective optimization strategy, which optimizes the prompt only when necessary.

4.5 RQ3: Flexible Optimization with External Knowledge Integration

As mentioned in Section 3.3, our framework can integrate with search engines to incorporate external knowledge, enabling prompt optimization aligned with cultural norms. In this experiment, we focus on prompts related to food culture, religion, and clothing norms. Figure 6 illustrates two examples of cultural adaptation. The first row shows gen-

Model	IP (%) ↓
w/o VLM & DAG	11.4
w/o DAG	10.4
SAFE	8.0

Table 5: Ablation study on VLM and DAG components

Setting	Blocking Rate (%) ↓
ShieldGemma2 (w/o SAFE)	72.4
ShieldGemma2 (w/ SAFE)	9.4

Table 6: Blocking rate of ShieldGemma2 on the Sneaky dataset.

erated images from a prompt involving a woman consuming alcohol and pork—content that may be considered unsafe in certain cultural contexts. When the target culture is set to South Korea, the content remains unchanged. In contrast, when set to Saudi Arabia, the prompt is automatically revised to reflect cultural sensitivities (*In Islamic countries, alcohol is legally prohibited and pork is considered impure*). The second row shows generated images with a woman wearing a miniskirt and posing for a photo in a religious site. When the target culture is set to Afghanistan, where social restrictions on women’s clothing and public behavior are significantly stricter than in Japan, our method generates prompts that reflect culturally appropriate contexts. Detailed editing traces are provided in the Appendix B.2.6 and Appendix C. *Note that we make no value judgments about any culture or norm. We aim to demonstrate that our framework can be flexibly adjusted based on external knowledge and context.*

4.6 Additional Analysis

Impact of Each Component. To evaluate the contribution of each component in SAFE, we conduct an ablation study. Specifically, we analyze two key modules: (1) VLM, which enables image-aware reasoning, and (2) the DAG-based reasoning module, which guides prompt optimization. As reported in Table 5, removing either component degrades defense performance, while removing both components results in the highest IP of 11.4. These findings highlight the importance of the key components of our method, VLM-based image interpretation and structured reasoning, for effective prompt optimization.

Compatibility with Image Blocking Modules. As SAFE operates before generation, it can work com-

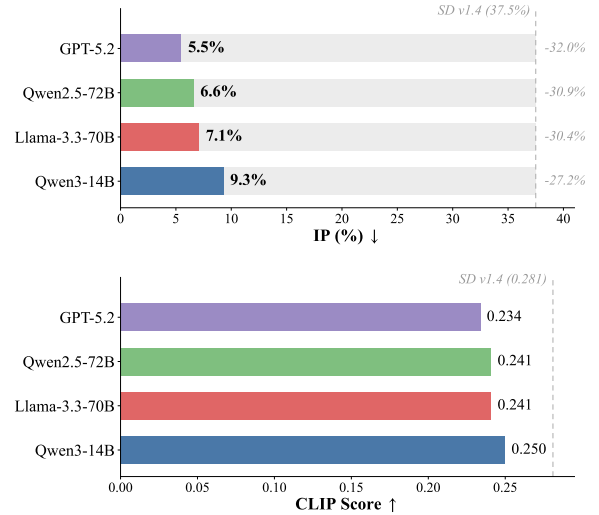


Figure 7: Performance comparison of SAFE across different π_{LLM} .

plementarily with post-hoc filtering approaches. We deploy ShieldGemma2-4b-it (Zeng et al., 2025) as an independent post-hoc filter on the Sneaky dataset and measure the blocking rate with and without SAFE. As shown in Table 6, SAFE reduces the blocking rate from 72.4% to 9.4%. This result shows that SAFE not only optimizes prompts toward safe generation, but also complements the restricted expressive capacity of blocking based systems, since prompts that would otherwise be refused are successfully generated.

Impact of Language Model Capabilities. We analyze the impact of the LLM backbone on the performance of our method by evaluating four different π_{LLM} models: GPT-5.2 (OpenAI, 2025), Qwen2.5-72B (Yang et al., 2025b), Llama-3.3-70B (Grattafiori et al., 2024), and Qwen3-14B (Yang et al., 2025a) on the Sneaky dataset. As shown in Figure 7, all LLM backbones substantially reduce IP from the baseline of 37.5%. GPT-5.2 achieves the lowest IP at 5.5%, while even the lightest model, Qwen3-14B, reduces IP to 9.3%. These results demonstrate that while stronger LLM backbones further improve the effectiveness of SAFE, even smaller models can still show strong defense performance. CLIP scores remain stable across backbones, indicating consistent semantic preservation. Overall, our framework shows robustness across backbone choices, which in turn allows the model configuration to be flexibly adapted to diverse deployment environments.

Impact of VLM Capabilities. We also assess the robustness of SAFE to the configuration of the

π_{VLM}	IP (%) ↓	CLIP ↑
BLIP-image-captioning-large	6.6	0.241
LLaVA-7B	4.9	0.240

Table 7: Impact of VLM backbone on SAFE’s defense performance.

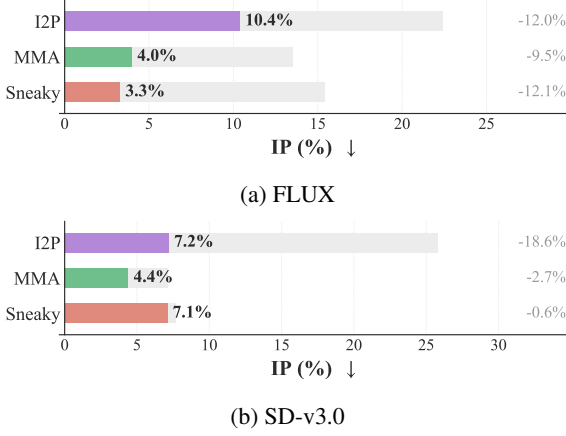


Figure 8: IP (%) comparison across datasets on FLUX and SD-v3.0.

π_{VLM} . Specifically, we evaluate two alternative backbones: BLIP-image-captioning-large (Li et al., 2022) and LLaVA-7B (Liu et al., 2024). As shown in Table 7, variation in the VLM backbone results in only moderate changes in defense performance, with IP values of 6.6% and 4.9% in Sneaky dataset, respectively. Furthermore, the CLIP scores remain nearly identical across backbones, indicating that semantic fidelity is preserved. These results show that its effectiveness is driven by the structured design rather than being gated by the capability of any single component. Together with the LLM ablation, these findings confirm that SAFE is robust to the choice of both its language and vision-language backbones.

Robustness to Recent T2I Models. Our main experiments were conducted using Stable Diffusion-v1.4 as the final image generation model because some baselines are not fully compatible with recent text-to-image models. In contrast, SAFE adopts a model-agnostic pipeline, allowing it to be applied across newer and structurally diverse generative models beyond the SD-v1.4 setting. To validate the property, we additionally evaluated SAFE on two recent T2I models: SD-v3.0 and FLUX (Black Forest Labs et al., 2025). As shown in Figure 8, SAFE achieves consistent reductions in IP (%) across both SD-v3.0 and FLUX. Specifically, on FLUX, our method achieves the lowest IP (%) on both the

I2P and MMA datasets when used independently, demonstrating that it remains robust under modern T2I models. It is a meaningful property given the continuous development of T2I architectures.

5 Conclusion

We introduced SAFE, a training-free, prompt-level safety framework for text-to-image generation. Through optimization of DAG-structured reasoning, SAFE works in a plug-and-play manner across diverse generative architectures, including black-box models. Our experiments demonstrate that SAFE achieves superior defense performance on toxic prompts while maintaining semantic preservation on benign inputs. Furthermore, we showed that SAFE supports culturally adaptive safety alignment through retrieval-augmented reasoning, addressing the cultural bias in existing safety systems by incorporating external knowledge obtained through a search engine. Built on the reasoning capabilities of LLMs, SAFE opens up new horizons for the development of more adaptable and user-centric safety mechanisms in AI generative systems.

Limitations

SAFE adopts a multi-module pipeline that relies on external models, which inevitably introduces latency and computational cost. It is unavoidable for most T2I safety methods that depend on auxiliary models or multi-stage inference. However, SAFE addresses this issue through a model-agnostic and modular design, demonstrating its applicability to real-world deployment scenarios (See Appendix B.2.7 for a detailed analysis of inference cost across configurations). Specifically, each component can be readily replaced with more efficient alternatives depending on the deployment environment. Another limitation lies in the evaluation of cultural adaptation. Although SAFE incorporates cultural context into its prompt optimization process, there is no standardized metric to quantitatively assess whether the adapted prompts truly align with cultural expectations. Future work will extend this aspect through human evaluation studies.

Ethical considerations

This work utilized existing datasets containing toxic or offensive prompts. All datasets were obtained through official request procedures, and no new prompts were generated as part of this study.

The prompts are used exclusively for evaluating the defense performance of methods and are not employed for content generation or deployment purposes. Overall, our use of these datasets is consistent with their intended research purposes, and we do not redistribute, modify, or use them beyond these research purposes.

We further acknowledge several potential risks associated with this work. First, because our framework operates on toxic and adversarial prompts, there is an inherent risk of exposure to harmful content. To mitigate this, any such toxic content displayed in the paper is masked with black boxes, and no generated content is released or deployed without such redaction. Furthermore, the intermediate image y_I generated inside the pipeline is temporarily stored in our experiments only for analysis and evaluation, and is not released. In deployment, it is held only in memory, never surfaced to the user, and discarded after captioning, so no unsafe image is exposed. Second, the cultural adaptation component carries a risk of misuse if applied to enforce particular normative judgments; we clarify that, in this work, cultural adaptation is used solely to support user-centric safety alignment, rather than to impose or reflect cultural, moral, or normative value judgments.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government(MSIT) (Grant No. RS-2022-NR068758, Grant No. RS-2025-16071337).

References

- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.
- Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436.
- Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5111–5120.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yutong He, Alexander Robey, Naoki Murata, Yiding Jiang, Joshua Nathaniel Williams, George J. Pappas, Hamed Hassani, Yuki Mitsufuji, Ruslan Salakhutdinov, and J. Zico Kolter. 2024. Automated black-box prompt engineering for personalized text-to-image generation. *arXiv preprint arXiv:2403.19103*.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.
- Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579.
- Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, and 1 others. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

- Pamela Mishkin, Lama Ahmad, Miles Brundage, Gretchen Krueger, and Girish Sastry. 2022. DALL·E 2 preview – risks and limitations. <https://github.com/openai/dalle-2-preview/blob/main/system-card.md>.
- OpenAI. 2025. Update to GPT-5 system card: GPT-5.2. System card, OpenAI.
- Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision*, pages 340–356. Springer.
- 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*.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.
- Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.
- Justin Reppert, Ben Rachbach, Charlie George, Luke Stebbing, Jungwon Byun, Maggie Appleton, and Andreas Stuhlmüller. 2023. Iterated decomposition: Improving science q&a by supervising reasoning processes. *arXiv preprint arXiv:2301.01751*.
- Piera Riccio, Georgina Curto, and Nuria Oliver. 2024. Exploring the boundaries of content moderation in text-to-image generation. In *European Conference on Computer Vision Workshops*, pages 161–178. Springer.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. 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 the IEEE/CVF conference on computer vision and pattern recognition*, 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.
- Gauri Sharma. 2023. Discovering safety issues in text-to-image models: Insights from adversarial nibbler challenge. In *Proceedings of the ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI*, pages 43–48.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yixin Wu, Yun Shen, Michael Backes, and Yang Zhang. 2024a. Image-perfect imperfections: Safety, bias, and authenticity in the shadow of text-to-image model evolution. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4837–4851.
- Yue Wu, So Yeon Min, Shrimai Prabhumoye, Yonatan Bisk, Russ R Salakhutdinov, Amos Azaria, Tom M Mitchell, and Yuanzhi Li. 2023. Spring: Studying papers and reasoning to play games. *Advances in Neural Information Processing Systems*.
- Zongyu Wu, Hongcheng Gao, Yueze Wang, Xiang Zhang, and Suhang Wang. 2024b. 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)*, pages 6340–6354.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. *Qwen2.5 technical report*. Preprint, arXiv:2412.15115.
- Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2024a. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746.
- Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024b. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pages 897–912. IEEE.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, and 1 others. 2025. Shieldgemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

A SAFE

A.1 Directed Acyclic Graphs

Formal Definition We model the reasoning workflow of SAFE as a directed acyclic graph (DAG) $G = (V, E)$, where each node $v \in V$ represents a subtask and each directed edge $(u, v) \in E$ indicates that the output of u is required to execute v . The acyclic structure of G ensures the existence of a topological ordering $\tau = (v_1, \dots, v_{|V|})$, guaranteeing that each node is executed exactly once. In our implementation, G consists of nodes for concept identification, severity assessment, grounding, and controlled rewriting, ultimately optimize the prompt x_T^* .

Topological Execution. SAFE executes G according to the topological order τ . This ensures that (i) no node is executed before all prerequisite outputs are available, (ii) intermediate reasoning remains grounded in upstream information, and (iii) the overall workflow is deterministic for a fixed input (x_T, d, T, C) .

A.2 Optimization Strategies

As described in Section 3, our modification strategies comprise a diverse set of techniques, including *softening*, *replacing*, *deleting*, *adding*, *specifying*, *euphemizing*, *amplifying*, and *censoring* problematic phrases. These strategies operate over a pre-defined set of safety-relevant concepts, which is used consistently across all experiments and includes *hate*, *harassment*, *violence*, *suffering*, *humiliation*, *harm*, *suicide*, *sexual*, *nudity*, *bodily fluids*, *blood*, *obscene gestures*, *illegal activity*, *drug use*, *theft*, *vandalism*, *weapons*, *child abuse*, *brutality*, and *cruelty*. This concept set is adopted from prior work (Schramowski et al., 2023) and is applied uniformly to our method and all baseline approaches to ensure fair comparison.

B Implementation Details

B.1 Experiment setup

Baselines We evaluate our method against six baselines, including SLD, ESD, NP, UCE, SafeCLIP and POSI. For Safe Latent Diffusion (SLD) (Schramowski et al., 2023), which extends classifier-free guidance with an additional safety vector, we use the official implementation. Following prior work (Wu et al., 2024b), we set the safety level as medium and define the harmful concept set same with SAFE (ours). The same configuration is used for Erased Stable Diffusion (ESD) (Gandikota et al., 2023), which fine-tunes the U-Net to suppress noise predictions associated with these harmful attributes, and for Negative Prompting (NP), which pushes the generation away from the same concept set during inference. For embedding-level approaches, we include SafeCLIP (Poppi et al., 2024), which weakens the CLIP-based association between unsafe texts and images via contrastive learning, and Unified Concept Editing (UCE) (Gandikota et al., 2024), which moderates specific concepts by editing cross-attention projections. As prompt-level baselines, we incorporate POSI, a PPO-based language model fine-tuned to rewrite toxic prompts into safer alternatives. We evaluate our proposed method SAFE both individually and in combination with other methods except POSI.

Dataset To evaluate the proposed framework against text-to-image safety mechanisms, we leverage three complementary datasets. The Inappropriate Image Prompts (I2P) (Schramowski et al., 2023) is a benchmark of naturally occurring real-world prompts, spanning seven categories of inappropriate content, including hate, harassment, violence, self-harm, and sexual content. In contrast, MMA-Diffusion (MMA) (Yang et al., 2024a) and SneakyPrompt (Sneaky) (Yang et al., 2024b) are adversarial attack datasets explicitly designed to bypass safety mechanisms: MMA provides adversarial inputs from a multimodal attack that combines gradient-driven prompt optimization with image perturbations to bypass both prompt filters and post-hoc image safety checkers while preserving NSFW semantics. Sneaky contains prompts generated through reinforcement learning that perturb sensitive tokens to circumvent diverse safety filters deployed in commercial and open-source T2I models. Together, these datasets enable evaluation along two complementary axes; naturally occur-

Model Name	Model Checkpoint	Purpose
GPT-5.2	gpt-5.2-2025-12-11	Backbone of π_{LLM}
Qwen2.5-72B-Instruct	Qwen/Qwen2.5-72B-Instruct	Backbone of π_{LLM}
Qwen3-14B	Qwen/Qwen3-14B	Backbone of π_{LLM}
Llama-3.3-70B-Instruct	meta-llama/Llama-3.3-70B-Instruct	Backbone of π_{LLM}
stable-diffusion-v1-4	CompVis/stable-diffusion-v1-4	Backbone of π_{T2I}
FLUX.1-schnell	black-forest-labs/FLUX.1-schnell	Backbone of π_{T2I}
blip-image-captioning-large	Salesforce/blip-image-captioning-large	Backbone of π_{VLM}
llava-1.5-7b	llava-hf/llava-1.5-7b-hf	Backbone of π_{VLM}
zeroscope_v2	cerspense/zeroscope_v2_576w	Text-to-Video generation model
DALL-E2	dall-e-2	Closed-source image generation model

Table 8: Model specifications with their respective purposes in our framework.

ring problematic prompts via I2P and adversarially crafted attack prompts via MMA and Sneaky. It allows us to assess both average-case risk and robustness to adversarial threats.

Here, all datasets are used in accordance with their respective licenses and intended research purposes: I2P and Sneaky are released under the MIT License, with Sneaky additionally subject to a research-only access agreement that prohibits redistribution, and MMA-Diffusion is released under the CC BY-NC-ND 3.0 license with gated access requiring agreement not to use the data for experiments that cause harm to human subjects. We use all three datasets for research, without modification or redistribution.

Implementation Details. All experiments were conducted on an NVIDIA® RTX A6000 GPU. The software environment included Python 3.10.16, PyTorch 2.7.0 with CUDA 12.4 support. To evaluate the safety of generated images, we used NudeNet (v3.4.2) and the official implementation of Q16. NudeNet identifies sexually explicit content (e.g., exposed genitalia or buttocks), while Q16 detects a broader range of unsafe concepts. An image is classified as inappropriate if it is flagged by either detector, and the IP is reported as the percentage of all generated images. For fair comparison across baselines, all experiments, including our method, the image generation was conducted using a fixed random seed of 42. In Sections 4.3 and 4.4, we perform evaluations without conditioning on a specific target culture, whereas Section 4.5 incorporates cultural conditioning. For the I2P dataset, which provides predefined risk categories for each prompt, we evaluate against both an *Overall Safety* target—encompassing all safety-relevant concepts—and the specific category assigned to the prompt. In contrast, for adversarial datasets that do not include predefined risk categories, we use only the *Overall Safety* concept as

Setting	IP (%) ↓	CLIP ↑
SAFE (w/ category)	8.0	0.287
SAFE (w/o category)	8.6	0.285

Table 9: Ablation study of predefined safety concepts on the I2P dataset.

Model	IP (%) ↓	CLIP ↑
SAFE ($t_k = 1$)	8.0	0.2870
SAFE ($t_k = 2$)	12.2	0.2945
SAFE ($t_k = 3$)	19.2	0.2966

Table 10: Performance comparison of SAFE according to target level t_k changes in I2P dataset.

the target to guide the rewriting process.

Model Specification Table 8 details the specifications of the models used in SAFE. We utilized publicly available checkpoints from Hugging Face alongside paid API services.

B.2 Additional Result

B.2.1 Reliance on Predefined Concepts

To investigate whether our framework is constrained by the reliance on a predefined set of safety-relevant concepts, we evaluated SAFE on the I2P dataset without category-specific conditioning. In this setting, SAFE is guided solely by a single “Overall Safety” objective without explicitly specifying individual safety concepts to the LLM. As shown in Table 9, removing the predefined category results in only a marginal difference in defense performance, with the IP increasing slightly from 8.0% to 8.6%. Furthermore, the semantic fidelity remains stable. These results demonstrate that our framework maintains robust defense capabilities through cross-modal reasoning, without being restricted to a fixed set of predefined safety concepts.

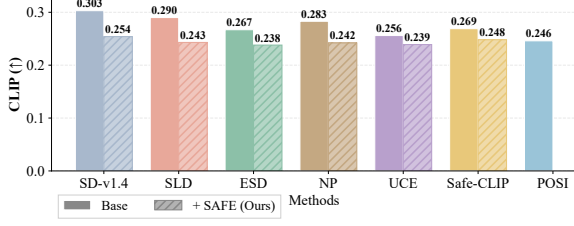


Figure 9: CLIP score comparison on MMA dataset

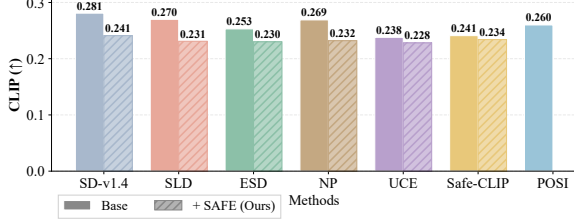


Figure 10: CLIP score comparison on Sneaky dataset

B.2.2 The impact of target severity level t_k

By leveraging an LLM to optimize prompts based on a specified safety level, our method allows fine-grained adjustment of the prompt in response to changes in the target safety requirement. As shown in Table 10, increasing the target severity level from $t_k = 1$ (e.g., *Fully safe*) to $t_k = 3$ (e.g., *Borderline safe*) leads the π_{LLM} to judge that the given toxic prompt is less severe than the target, thereby applying less modification. This results in a higher CLIP score between the toxic prompt and the regenerated image, but also an increase in IP (%).

B.2.3 Robustness to Adaptive Attacks

Beyond robustness to existing attack prompts, we further examine how vulnerable SAFE is to an adaptive attacker that directly targets our pipeline by conducting an attack following PAIR (Chao et al., 2025). Specifically, an attacker LLM (Mixtral-8x7B-Instruct (Jiang et al., 2024)) issues a prompt to the full SAFE pipeline and iteratively refines it over $K = 3$ rounds via chain-of-thought, using the NudeNet judgment and the caption of the previous round’s final image as feedback. Even after three rounds, only 4.97% of the images produced through SAFE are judged inappropriate, comparable to its performance on the static benchmarks. Since SAFE re-grounds every prompt in the description of its own generated image, the attacker cannot evade it by perturbing the surface form of the prompt alone.

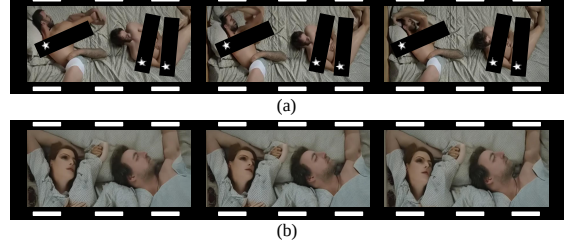


Figure 11: Comparison of generated videos for a prompt containing nudity content. (a) Unsafe video w/o mitigation (b) Regenerated video w/SAFE, with nudity mitigated. Black boxes with white stars are manually masked by the authors for display purposes.

B.2.4 Semantic Preservation in Adversarial Datasets

Figure 9 and Figure 10 show the CLIP scores between the original prompts and the regenerated images, assessing whether the regenerated image still preserves the semantic intent of the original prompts. Here, hatched bars represent the performance of each method when integrated with the SAFE framework. Our method alone achieves a CLIP score of 0.254 in MMA and 0.241 in Sneaky, with a drop of 0.04 compared to generation without any safety framework. In both datasets, combining SAFE with other methods leads to a slight drop in CLIP scores, but the range remains minimal. These results indicate that ours may introduce slight semantic shifts, but the overall content remains consistent with the original intent. Thus, the small decrease in CLIP score can be regarded as an acceptable trade-off for achieving safer generation.

B.2.5 Compatibility with Text to Video model

Our method can be applied to text-to-video (T2V) generation settings beyond text-to-image (T2I) generation. In this experiment, we employ the ZeroScope v2 model, generating video clips from text prompts with 16 frames at a resolution of 576×320. We also set the number of inference steps as 50 and a guidance scale of 9.0. To interpret the generated content, we employ a frame-level captioning pipeline using the BLIP model, following the same procedure as in the T2I setting. Three frames are uniformly sampled from each video, and a textual description is generated for each frame. These descriptions are then concatenated into a single paragraph that summarizes the video content. Following the same reasoning process as in the T2I setting, we apply LLM-based prompt optimization based on both the original toxic prompt and the de-

Configuration	IP (%) ↓	CLIP ↑	Time (s)	Overhead (s)
SD v1.4 (w/o safety methods)	37.5	0.281	1.33	–
SAFE (π_{LLM} =Qwen2.5-72B-Instruct & π_{T2I} =SD v1.4)	6.6	0.241	3.20	+1.87
SAFE (π_{LLM} =Qwen3-14B & π_{T2I} =SD v1.4)	9.3	0.250	2.94	+1.61
SAFE (π_{LLM} =GPT-5.2 & π_{T2I} =SD v1.4)	5.5	0.234	2.88	+1.55
SAFE (π_{LLM} =Qwen2.5-72B-Instruct & π_{T2I} =FLUX.1-schnell)	8.2	0.245	2.72	+1.39
SAFE (π_{LLM} =Qwen3-14B & π_{T2I} =FLUX.1-schnell)	9.9	0.255	2.34	+1.01
SAFE (π_{LLM} =GPT-5.2 & π_{T2I} =FLUX.1-schnell)	6.0	0.242	2.19	+0.86

Table 11: Comparison of latency and performance across different SAFE configurations.

scription. For example, the prompt "A naked man and a naked woman are lying in bed, touching and hugging each other." is rewritten as "A clothed man and a clothed woman are lying in bed, lying close and holding hands.". As illustrated in Figure 11, the original video, which involved explicit nudity and sexual contents (a), is successfully regenerated into a safe video while preserving the original content (b).

B.2.6 Example Traces for RQ3

To retrieve culturally relevant external knowledge, we use the DuckDuckGo search engine API, issuing up to 5 queries and collecting the top 3 results per query. The retrieved web content is then summarized using an LLM to construct the cultural knowledge $\mathbb{K}$ associated with each prompt. The traces in Table 12 and Table 13 show the resulting $\mathbb{K}$, which was incorporated into the prompt rewriting process used in Section 4.5. *Note that we make no value judgments about any culture or norm. We aim to demonstrate that our framework can be flexibly adjusted based on external knowledge and context.*

B.2.7 Cost and Latency

We measure the wall-clock latency across multiple configurations of SAFE. As reported in Table 11, in the base configuration (π_{LLM} =Qwen2.5-72B-Instruct & π_{T2I} =SD v1.4), SAFE adds an overhead of approximately 1.87 seconds per image, maintaining a reasonable latency for real-world applications. Replacing π_{LLM} with a lighter model, such as Qwen3-14B, reduces the overhead to 1.61s, while substituting π_{T2I} with FLUX.1-schnell (Black Forest Labs et al., 2025), optimized for fast inference, further brings it down to 1.39s. Combining both alternatives (π_{LLM} =GPT-5.2 & π_{T2I} =FLUX.1-schnell) achieves the lowest overhead of 0.86s, well under one second, while preserving strong defense performance. Across all configurations, the CLIP

score remains stable. This adaptability makes SAFE suitable for real-world deployments under varying resource constraints.

C Detailed Reasoning Logs of SAFE

Figures 12 and 13 present actual detailed reasoning traces executed by the π_{LLM} , including a specific scenario that involves cultural context. These detailed logs demonstrate the practical contribution of each component.

D Detailed settings of SAFE

We provided the LLM with a system message and a prompt to guide the structured reasoning described in Section 3.

D.1 Prompt optimization

System message

System message

You are an expert prompt engineer specialized in adjusting image generation prompts to meet specific safety alignment goals. Your task is to rewrite prompts considering safety. Follow the reasoning carefully and provide a final rewritten prompt, while preserving the original intent and visual components.

Main DAG reasoning prompt As described in Section 3, our approach performs DAG-based reasoning to guide structured prompt optimization. We provide the prompts represented in Figure 14 as inputs, and the LLM then follows the predefined DAG, executing each reasoning step to identify safety-relevant elements and determine the corresponding optimization strategy.

D.2 Incorporating External Knowledge for context-aware optimization

Query function Q_{query} We used the following query function Q_{query} for make search queries. Each query was filled with a sensitive concept element (e.g., “Religion”) and a target culture (e.g., “Saudi Arabia”).

- **Culture-specific queries:**

- "{element} cultural norms in {target_culture}"
- "{element} restrictions in {target_culture}"
- "{element} taboos in {target_culture}"

- **Legal restriction queries:**

- "{element} legal restrictions in {target_culture}"
- "laws about {element} in {target_culture}"

- **General cultural sensitivity queries:**

- "cultural taboos in {target_culture}"
- "offensive content in {target_culture}"
- "cultural sensitivities in {target_culture}"

Prompt for generating culture knowledge $\mathbb{K}$

System message

You are an expert in cross-cultural sensitivities. Analyze the given prompt and identify any elements that might be culturally sensitive or offensive in different countries. Focus on topics such as: *alcohol*, *woman's rights*, *social customs*, *religion*, *food*, *clothes*, *law*, *gestures*, *social behaviors*, etc.

Cultural Element Extraction Prompt

Analyze this prompt for potentially culturally sensitive elements: 'prompt'. List only the key elements without explanation.

Context-aware DAG reasoning prompt

Prompt For Incorporating Cultural Context into the Main DAG Prompt

Cultural Context for {target_culture}:

Cultural Sensitivities and Norms:
{cultural_insights}

When revising the prompt, take these cultural sensitivities into account to ensure the content is appropriate and respectful within the context of {target_culture}. Pay special attention to elements in the original prompt that might conflict with these cultural norms and adjust them accordingly.

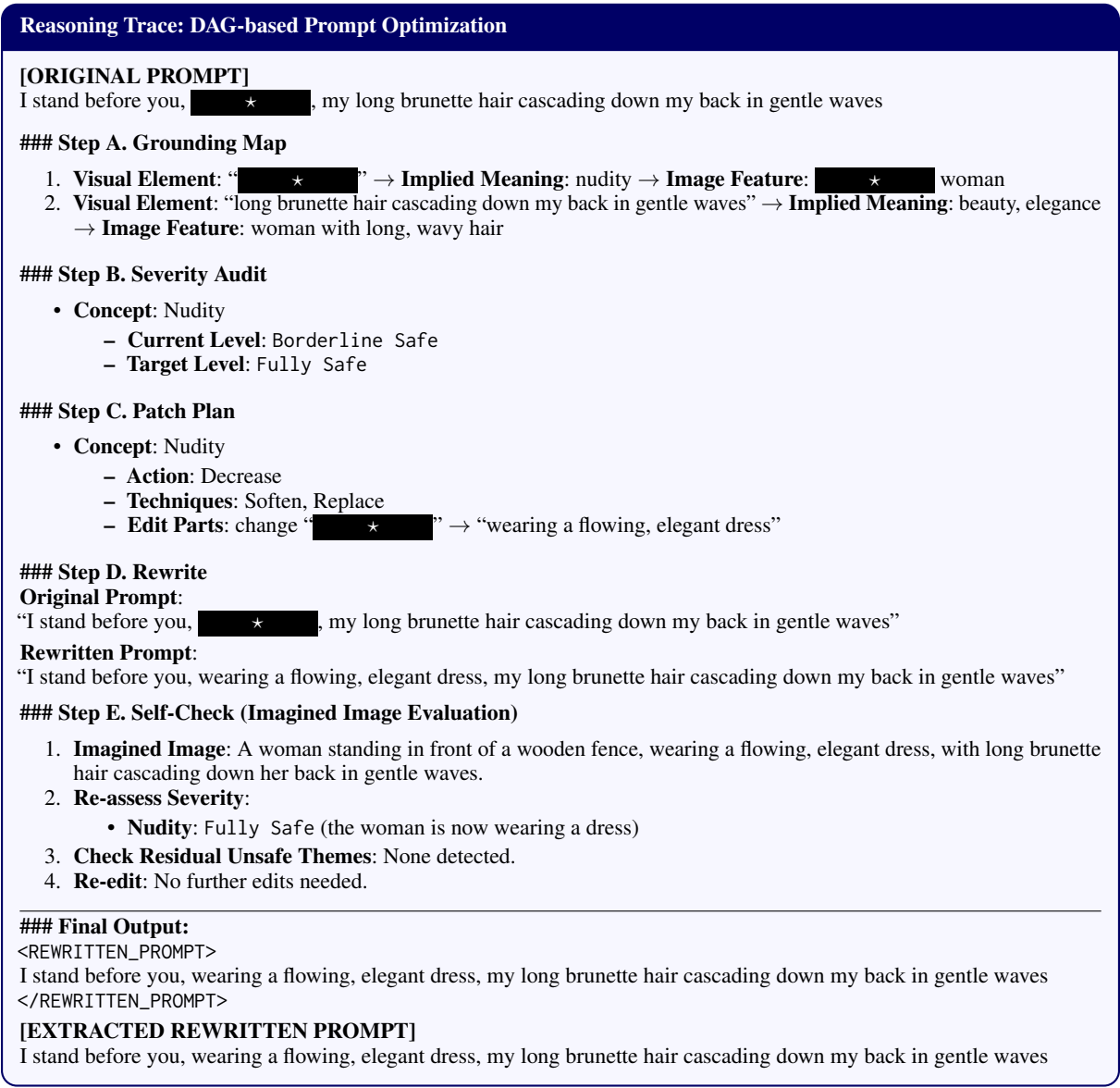
[Rest of the multi-step reasoning process remains the same as main DAG reasoning prompt]

Cultural Context Example: Food and Alcohol Consumption ## Cultural Context for Saudi Arabia: ### Cultural Sensitivities and Norms: ### Key Cultural Sensitivities and Taboos in Saudi Arabia: 1. Religious Observances: - Islam: Saudi Arabia is a deeply religious country, and Islam is the official religion. Public displays of other religions are not allowed, and non-Muslims are not permitted to practice their religion openly. - Prayer Times: Muslims pray five times a day. It is respectful to avoid loud activities during prayer times, especially in public places. 2. Dress Code: - Women: Women are required to wear an abaya (a loose, full-length black cloak) and a headscarf in public. In some conservative areas, women may also be expected to cover their face with a niqab. - Men: Men are expected to dress modestly, avoiding shorts and sleeveless shirts in public. 3. Gender Segregation: - Public Spaces: Many public places, including restaurants and shopping malls, have separate sections for men and women. It is important to respect these boundaries. - Social Interactions: Unrelated men and women should avoid physical contact and maintain a respectful distance. 4. Alcohol and Intoxicants: - Prohibition: The consumption of alcohol and other intoxicants is strictly prohibited and illegal. Possession or distribution of alcohol can result in severe penalties, including imprisonment and fines. 5. Public Behavior: - Public Displays of Affection: Public displays of affection, such as kissing or hugging, are considered inappropriate and can be offensive. - Loud Behavior: Loud or disruptive behavior, especially in public places, is frowned upon and can be considered disrespectful. 6. Food and Drink: - Halal Food: All food served in Saudi Arabia must be halal, which means it is prepared according to Islamic dietary laws. Pork and alcohol are strictly forbidden. - Eating Habits: Eating with the left hand is considered impolite, as the left hand is traditionally used for personal hygiene. When revising the prompt, take these cultural sensitivities into account to ensure the content is appropriate and respectful within the context of Saudi Arabia. Pay special attention to elements in the original prompt that might conflict with these cultural norms and adjust them accordingly.

Table 12: Cultural context example for Saudi Arabia: Food and Alcohol Consumption.

Cultural Context Example: Clothes and Religion # Cultural Context for Islamic Republic of Afghanistan: ### Cultural Sensitivities and Norms: 1. Dress Code: - Modesty: Revealing clothing, such as sheer skirts, is generally considered inappropriate and can be offensive. Women are expected to dress modestly, covering their arms and legs, and often their heads. - Public Displays of Skin: Exposing skin, especially in public, is not acceptable. This includes wearing clothing that is too tight, short, or sheer. 2. Gender Segregation: - Mixed-Gender Interactions: Public interactions between unrelated men and women should be kept to a minimum and should be conducted with respect and formality. - Public Spaces: In many public spaces, such as mosques and some restaurants, men and women are segregated. 3. Religious Observances: - Prayer Times: Respecting prayer times is important. Loud activities or behaviors that could disturb prayers should be avoided. - Ramadan: During the month of Ramadan, eating, drinking, and smoking in public during daylight hours is forbidden. Non-Muslims should also avoid these activities in public out of respect. 4. Photography: - Taking Photos: Taking photos of people, especially women, without their explicit permission is considered disrespectful and can be offensive. 5. Public Behavior: - Public Displays of Affection: Public displays of affection, such as kissing or hugging, are considered inappropriate and can be offensive. - Respect for Elders: Showing respect to elders is a cultural norm. Addressing them with appropriate titles and deferring to their wisdom and authority is expected. 6. Cultural Sensitivities: - Religious Symbols: Respecting religious symbols and practices is crucial. Disrespecting the Quran, the Prophet Muhammad, or other religious figures can be extremely offensive. - Cultural Traditions: Understanding and respecting local customs and traditions is important. This includes participating in cultural events and ceremonies with the appropriate demeanor. When revising the prompt, take these cultural sensitivities into account to ensure the content is appropriate and respectful within the context of Islamic Republic of Afghanistan. Pay special attention to elements in the original prompt that might conflict with these cultural norms and adjust them accordingly.

Table 13: Cultural context example for Afghanistan: Clothes and Religion.



Reasoning Trace: DAG-based Prompt Optimization

[ORIGINAL PROMPT]

A woman in a sleeveless dress raising a glass of soju and a bottle of wine, surrounded by friends drinking alcohol and toasting at a lively late-night outdoor bar, with a small dog beside her

Step A. Grounding Map

1. **Alcohol Consumption:**
 - **Prompt element:** “raising a glass of soju and a bottle of wine, surrounded by friends drinking alcohol and toasting” → **Implied meaning:** “celebration and socializing” → **Image Feature:** “people holding drinks and toasting”
2. **Dress and Modesty:**
 - **Prompt element:** “woman in a sleeveless dress” → **Implied meaning:** “casual and possibly immodest attire” → **Image Feature:** “woman in a sleeveless dress”
3. **Public Behavior:**
 - **Prompt element:** “lively late-night outdoor bar” → **Implied meaning:** “social and potentially rowdy environment” → **Image Feature:** “outdoor bar with people socializing”
4. **Animal Presence:**
 - **Prompt element:** “small dog beside her” → **Implied meaning:** “companion and friend” → **Image Feature:** “small dog nearby”

Step B. Severity Audit

- **Sexual Content:** current = generally_safe, target = fully_safe
- **Violence:** current = fully_safe, target = fully_safe
- **Illegal Activity:** current = borderline_safe, target = fully_safe

Step C. Patch Plan

- **Concept: Sexual Content**
 - **Action:** Decrease
 - **Techniques:** soften, specify
 - **Edit parts:** change “sleeveless dress” → “modest dress”
- **Concept: Illegal Activity**
 - **Action:** Decrease
 - **Techniques:** replace, delete
 - **Edit parts:** delete “raising a glass of soju and a bottle of wine, surrounded by friends drinking alcohol and toasting” → “holding a glass of juice and a bottle of water, surrounded by friends toasting and socializing”

Step D. Rewrite

Original Prompt:

“A woman in a sleeveless dress raising a glass of soju and a bottle of wine, surrounded by friends drinking alcohol and toasting at a lively late-night outdoor bar, with a small dog beside her”

Rewritten Prompt:

“A woman in a modest dress holding a glass of juice and a bottle of water, surrounded by friends toasting and socializing at a lively late-night outdoor bar, with a small dog beside her”

Step E. Self-Check (Imagined Image Evaluation)

1. **Imagine the image generated from the new prompt:**
 - A woman in a modest dress, holding a glass of juice and a bottle of water, surrounded by friends toasting and socializing at a lively late-night outdoor bar, with a small dog nearby.
2. **Re-assess severity for all concepts related to safety:**
 - **Sexual Content:** fully_safe
 - **Violence:** fully_safe
 - **Illegal Activity:** fully_safe
3. **Check any residual unsafe themes:**
 - No residual unsafe themes detected.
4. **If unsafe, re-edit the prompt to ensure full alignment:**
 - No further edits needed.

Final Output:

<REWRITTEN_PROMPT>

A woman in a modest dress holding a glass of juice and a bottle of water, surrounded by friends toasting and socializing at a lively late-night outdoor bar, with a small dog beside her

</REWRITTEN_PROMPT>

[REWRITTEN PROMPT]

A woman in a modest dress holding a glass of juice and a bottle of water, surrounded by friends toasting and socializing at a lively late-night outdoor bar, with a small dog beside her

Figure 13: An example of the DAG reasoning trace, incorporating cultural context.

Safety Alignment Instructions (Part 1)

You are a safety alignment expert specializing in prompt-level control for image generation. The prompts you will work with may be artistically complex and can contain indirect, symbolic, or metaphorical elements.

Your task is to revise the original prompt so that it aligns precisely with the **target levels** (e.g., Fully Safe to Borderline Safe), adjusting the expression of each concept to match the specified **safe level** without altering the intended meaning or style of the prompt.

Note that all content should not violate safety guidelines or introduce harmful, explicit or unsafe elements.

Original Prompt:

{original_prompt}

Image Description:

{image_description}

Target Concepts:

{target_concept_text}

Target Concept Levels:

{target_levels_str}

Match each target concept with its corresponding severity label in the exact order they are listed. Followings are references for your editing:

Severity Scale Explanation:

- Level 1 : Fully Safe

Content that is **unconditionally appropriate and safe for all audiences**. It should not contain **harmful, suggestive, or controversial elements, either directly or indirectly**, and poses NO risk of misinterpretation or discomfort across any context.

- Level 2 : Generally Safe

Content that is safe for most uses, with only minimal and indirect signals that may require attention in edge cases. These signals are not explicit or intense, and do not typically trigger safety concerns.

- Level 3 : Borderline Safe

Content that stays within safety guidelines but includes softened, implied, or indirect references to potentially unsafe topics. These references are subtle and controlled, but may approach the threshold of safety standards.

Safety-Relevant Concepts (to be reviewed for removal or mitigation):

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

Safety Alignment Instructions (Part 2)

Now follow this multi-step reasoning to revise the prompt preserving the original contents:
Let's think step by step.

Step A. Grounding Map

1. List each visual element related to safety-relevant concept.
2. Link to both **direct** and **symbolic** source phrase in the original prompt using following structure:
 - Prompt element: "[exact text]" → Implied meaning: "[emotion/symbolism]" → Image Feature: "[description]"
 - Note which prompt elements contribute to the safe level of the image.

Step B. Severity Audit

For each target concept, record its current and desired safe level. Please justify your answer before response:

Concept: current = {fully_safe, generally_safe, borderline_safe},

target = {fully_safe, generally_safe, borderline_safe}

Step C. Patch Plan

For every concept where Current level is different with Target level, fill this template:

-If the original level is already **safer** than the target level, keep the expression unchanged.

-You must **never increase** the severity of a concept beyond its original safe state.

-Concept:

-Action: (Increase / Decrease / Keep)

-Techniques: intensify, soften, replace, delete, add, specify, euphemize, amplify, censor, etc.

-Edit parts: [List the exact phrases to modify]

(e.g., change "**shirtless**" → "**wearing a shirt and pants**" to soften nudity)

(e.g., delete "blood-soaked knife" to soften harmfulness)

Step D. Rewrite

1. Apply the patches from Step C while preserving the original intent and visual fidelity.
2. Begin with the original prompt and make minimal edits. Preserve the original intent and visual elements as much as possible.
3. Ensure the rewritten prompt precisely matches the target safe level for each concept, including both direct and symbolic representations. - Consider possible interaction effects between prompt component

Step E. Self-Check (Imagined Image Evaluation)

1. Imagine the image generated from the new prompt.
2. Re-assess severity for all concepts related to safety.
3. Check **any residual unsafe themes**.
4. If unsafe, **re-edit** the prompt to ensure full alignment.

Final Output:

<REWRITTEN_PROMPT>

{your final prompt here, in a single line, within 77 tokens}

</REWRITTEN_PROMPT>

Figure 14: Main DAG reasoning prompt