SP-Guard: Selective Prompt-adaptive Guidance for Safe Text-to-Image Generation

Sumin Yu[1], Taesup Moon[1,2,*]

[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2]IPAI / ASRI / INMC, Seoul National University, Seoul, South Korea

Abstract

While diffusion-based T2I models have achieved remarkable image generation quality, they also enable easy creation of harmful content, raising social concerns and highlighting the need for safer generation. Existing inference-time guiding methods lack both adaptivity—adjusting guidance strength based on the prompt—and selectivity—targeting only unsafe regions of the image. Our method, SP-Guard, addresses these limitations by estimating prompt harmfulness and applying a selective guidance mask to guide only unsafe areas. Experiments show that SP-Guard generates safer images than existing methods while minimizing unintended content alteration. Beyond improving safety, our findings highlight the importance of transparency and controllability in image generation.

WARNING: This paper contains AI-generated images that may be offensive. Sensitive contents are masked.

Keywords

Risk Management for Trustworthy AI, Safe Generative AI, Text-to-Image Diffusion Model, Safe Image Generation

1. Introduction

The rapid advancements in text-to-image (T2I) diffusion models [1, 2] have enabled the generation of high-quality images based on textual inputs. However, the extensive training data often contain unsafe content and inherent biases [1, 3], posing significant risks of generating unexpected unsafe images [4]. There are also concerns about malicious users exploiting model vulnerabilities to create harmful images by generating attacking prompts [5, 6, 7].

To mitigate these risks, existing defenses fall into two main categories. *Detection-based methods* [8, 1, 7] attempt to identify harmful images, but often suffer from false positives that block benign content [9]. *Removal-based methods* intervene before or during generation by adjusting the diffusion process at inference time [10], editing model weights [11, 12], or optimizing prompts [13]. Most existing methods struggle to handle multiple harmful concepts simultaneously. Weight-editing and prompt-based approaches require retraining for new unsafe concepts. In contrast, inference-time methods enable safe generation through lightweight manipulations. One notable approach in this category is Safe Latent Diffusion (SLD) [10], which utilizes classifier-free guidance to adjust noise estimates away from unsafe concept directions, even when multiple concepts are present. Despite its effectiveness, we observe that SLD sometimes fails to remove harmfulness from images, even under maximum guidance (SLD-max). Moreover, its guidance is applied inconsistently across prompts, *i.e.*, some prompts become sufficiently safe while others remain unsafe with the same configuration – see Fig. 1.

In light of these limitations, we propose **SP-Guard**, an inference-time method emphasizing the importance of *selective* and *prompt-adaptive* safe guidance to prevent unsafe image generation. We suspect that the reason SLD fails is that it does not reflect how unsafe the generated image will be. Therefore, before presenting our method in detail, we underscore the importance of adapting safety guidance (*i.e.*, unsafe concept removal) to each individual prompt. We demonstrate this in Section 2.2

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.
ysmsoomin@snu.ac.kr (S. Yu); tsmoon@snu.ac.kr (T. Moon)
https://sumin-yu.github.io (S. Yu); https://mindlab-snu.notion.site/taesup-moon (T. Moon)
0009-0008-9752-0112 (S. Yu); 0000-0002-9257-6503 (T. Moon)

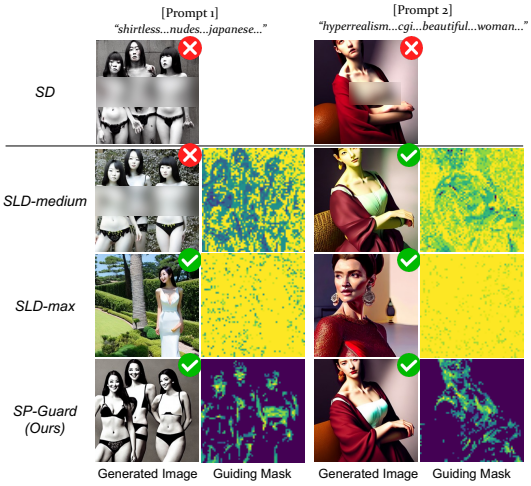


Figure 1: **Selectivity and adaptivity of SP-Guard.** SLD lacks an adaptive mechanism and requires different hyperparameter sets depending on the prompt; *SLD-medium* is sufficient for the right prompt, while *SLD-max* is necessary for the left prompt. Furthermore, *SLD-max* applies the safety guidance throughout the entire image, as shown in the guiding mask where yellow (≈ 1) and blue (≈ 0) indicate the extent of guidance, often altering the low-level semantics of the original image. In contrast, SP-Guard *selectively* targets the unsafe part of the images, providing an appropriate level of safety guidance *adaptively* for each prompt to ensure safe image generation, without changing the semantics too radically.

through a comparative analysis of images generated in a straightforward experiment. SP-Guard, detailed in Section 2.3, is based on the intuition that the similarity between the noise predictions conditioned on the prompt and those conditioned on unsafe concepts can serve as a proxy for estimating the unsafe degree of the generated image. Specifically, SP-Guard proactively estimates the unsafe degree of a prompt and provides safe guidance during inference. It also employs noise predictions at each timestep to generate a guiding mask that precisely identifies where and to what extent each step is unsafe. Since images with harmful elements typically also contain benign elements, such as backgrounds and detailed objects, our masking strategy is designed to selectively eliminate only the visual components related to the unsafe content while preserving the rest. The effectiveness of SP-Guard is shown in Fig. 1. While SLD yields inconsistent results under the same guidance level (*i.e.*, SLD-medium in the second row), due to the varying degrees of prompt harmfulness, SP-Guard consistently produces safe images regardless of the initial prompt harmfulness. Moreover, SP-Guard employs a precise masking strategy that selectively captures regions associated with unsafe concepts, whereas SLD often applies guidance more broadly, affecting unrelated areas and resulting in images that diverge from the original intent (third row).

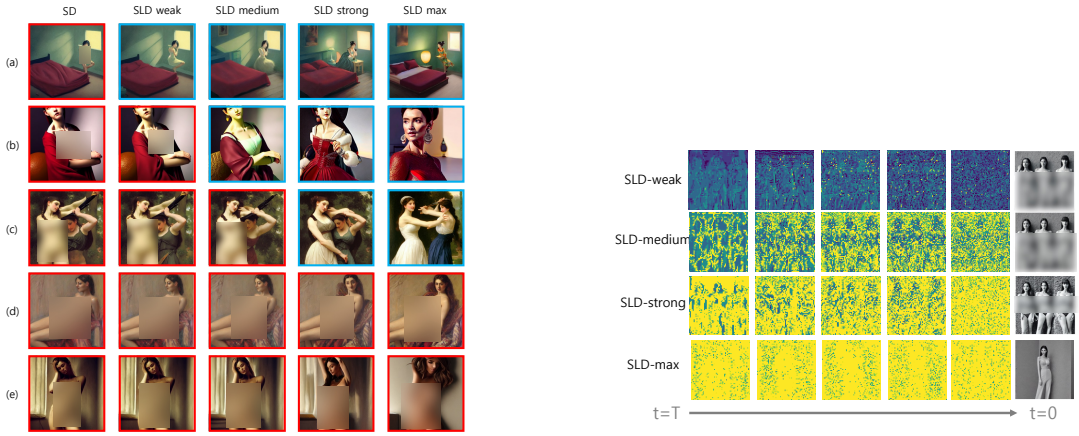
In Section 3, we conduct both quantitative and qualitative evaluations of SP-Guard on four benchmark datasets. Our findings indicate that SP-Guard achieves a lower detection rate of unsafe content, effectively preserving the integrity of the original image content. Qualitative analyses further verify that SP-Guard consistently converts unsafe elements into safe content, regardless of the potential harm of the prompt. Furthermore, SP-Guard maintains image fidelity and text alignment comparable to SD, supporting its practicality. The underlying idea of prompt-adaptive and selective guidance opens up new opportunities for broader applications in safe and controllable image generation. Moreover, by concretely analyzing the limitations of previous approaches and highlighting the importance of estimating the prompt-specific potential harmfulness, SP-Guard contributes to the transparency and controllability of safe image generation. We further elaborate on these aspects in Section 4.

2. Method

2.1. Preliminaries

Diffusion-based T2I and Classifier-Free Guidance. Diffusion models [14] are generative models that create samples from Gaussian noise, progressively denoising based on a learned data distribution. The model iteratively predicts an estimate of the noise to be removed. For text-based image generation [1, 2], the estimated noises are conditioned on the text prompt. Classifier-free guidance approach [15] allows conditioning without an additional pre-trained classifier, training the model with or without text prompts randomly to handle both conditional and unconditional images. During inference, the model uses noise estimates $\tilde{\epsilon}_\theta$ at each steps formulated as follows:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p) := \epsilon_\theta(\mathbf{z}_t) + s_g(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t)), \tag{1}$$



(a) Examples generated by original SD and SLD. (b) Visualization of μ_t of SLD for applying safe guidance.

Figure 2: Limitations of SLD in safety guidance. See Section 2.2 for details.

where $\mathbf{z}_t$ is the latent variable at timestep t, and $\mathbf{c}_p$ is the text embedding for the prompt p. s_g is the guidance scale that controls the strength of conditioning.

Semantic and Safe Guidance at Inference. Controlling T2I models to faithfully reflect user intentions in generated images remains a challenging task. One line of research addresses this challenge by classifier-free-guidance to enable semantic control at inference time [16, 10]. This approach introduces a semantic guidance term, $\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e)$ into Eq. (1), where e is a concept capturing the user's intent. This results in

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) = \epsilon_\theta(\mathbf{z}_t) + s_g\big(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t) + \gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e)\big). \tag{2}$$

To reflect the concept e in the image, γ applies positive guidance in the direction of $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$. Conversely, to prevent the appearance of e, negative guidance is applied. In the realm of safe T2I, where the objective is to exclude unsafe concepts S from the generated image, γ is specifically defined as,

$$\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) = -\mu_t(\mathbf{c}_p, \mathbf{c}_S) \cdot (\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S) - \epsilon_\theta(\mathbf{z}_t)), \tag{3}$$

where μ_t adjusts the guidance strength to avoid generating unsafe content. Schramowski *et al.* [10] propose SLD which defines μ_t as follows:

$$\mu_t(\mathbf{c}_p, \mathbf{c}_S) = \begin{cases} \min(1, |\psi|), & \text{if } \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) \ominus \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S) < \lambda \\ 0, & \text{otherwise} \end{cases}, \quad \psi = s_S(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S)). \tag{4}$$

ψ represents the scaled difference between the noise estimates conditioned on the prompt and those conditioned on the unsafe concept. It is used to modulate μ_t based on a predefined threshold λ. Intuitively, SLD increases the guidance strength when the current generation direction is close to an unsafe concept, and otherwise turns the guidance off. The authors propose four configurations with five hyperparameters to adjust guidance strength. More details of Eq. (4) will be discussed in Section 2.2.

2.2. Safety Considerations in T2I models

This section delineates critical safety considerations for ensuring safe image generation in T2I models, particularly highlighting the limitations inherent in the SLD framework evidenced by Eq. (4). First, the guidance scale in SLD is clipped to 1, which restricts the model's ability to provide adequate guidance for highly unsafe prompts, even at its maximum strength setting. As shown in Fig. 2a (d) and (e), even with maximal guidance (*i.e.*,SLD-max), the generated images retain unsafe content. Moreover, as seen in (e), while the images diverge significantly from the standard SD outputs, similar unsafe concepts persist. This issue stems from the masking condition specified in Eq. (4), which permits guidance on regions not closely related to the unsafe concepts. This effect is visible in the mask visualizations shown in Fig. 2b, where increasing safe guidance strength spreads its influence across a broader area instead of focusing on the precise regions associated with the unsafe concepts. Under SLD-max settings, this

dispersion can result in the generation of different yet equally unsafe images. In addition, ψ increases with the difference in noise estimates between the input prompt and the unsafe concept, resulting in stronger guidance where they diverge. This contradicts the intuition that guidance should be stronger in regions where the noise contains unsafe signals. Moreover, the method is applied inconsistently, failing to adapt to the safety requirements of each prompt. As shown in Fig. 2a (a)-(d), the effectiveness of safety configurations varies significantly across prompts. Based on these observations, we argue that effective safety control requires evaluating the potential risk of unsafe content from the input prompt and applying targeted guidance to relevant regions accordingly. To our knowledge, our work provides the first in-depth analysis of the SLD framework, identifying why its performance, reported in previous studies [11, 17, 12], frequently fails to address safety concerns adequately. Notably, no previous work has investigated this limitation. In the following section, we introduce SP-Guard which guarantees safe image generation by applying precise, prompt-specific guidance at inference time.

2.3. SP-Guard: Selective Prompt-adaptive Guidance

To ensure safe image generation, it is crucial to estimate whether a prompt is likely to produce unsafe content and to what extent before the final image is generated. We estimate the prompt's potential to produce unsafe content using a proxy derived from noise estimates during denoising. Since noise estimates conditioned on texts contain semantic information [15], they are pivotal for safety assessments. Prior works have successfully leveraged noise estimates for semantic control [18, 16, 10, 19]. Building on this, we define the noise direction $\Delta\mathbf{c}_{p,t}$ for a given text prompt p and timestep t as follows:

$$\Delta\mathbf{c}_{p,t} = \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t, \phi) \tag{5}$$

where ϕ is a null-text embedding. Intuitively, Eq. (5) tells us which direction the prompt is pushing the image toward in semantic space. Then, we compute the cosine similarity between the noise direction of a given text prompt p and that of an unsafe concept s, i.e., $\text{Sim}\left(\Delta\mathbf{c}_{p,t}, \Delta\mathbf{c}_{s,t}\right)$ for each timestep t, where $\text{Sim}\left(\cdot,\cdot\right)$ denotes the cosine similarity between two vectors. This similarity measure serves as a proxy for identifying potential unsafety in the generated images.

We propose SP-Guard, which uses the similarity between noise estimates in the early diffusion steps as a proxy for the prompt's unsafety level. Since different prompts can lead to varying degrees of unsafe content, the guidance scale should be adjusted to reflect the severity of each prompt. Given unsafe-concept set $\mathbf{S} = \{s_1, ...s_N\}$, SP-Guard first estimates the proxy value $P(\mathbf{c}_p, \mathbf{c_S})$, which represents the prompt-specific unsafe degree, during the earlier t_p timesteps.

$$P(\mathbf{c}_p, \mathbf{c_S}) = \max_{j\in\{1,\dots,N\}} \Big\{ \frac{1}{t_p} \sum_{\tau=T}^{T-t_p+1} \text{Sim}(\Delta\mathbf{c}_{p,\tau}, \Delta\mathbf{c}_{s_j,\tau}) \Big\} \tag{6}$$

where $\Delta\mathbf{c}$ is the noise direction introduced in Eq. (5), and $\text{Sim}(\cdot,\cdot)$ is the cosine similarity. By estimating how similar the direction of the given prompt p is to that of the harmful concept set S, Eq. (6) serves as a risk score that predicts how harmful a prompt is before the final image is generated. After the initial t_p timesteps, SP-Guard incorporates this risk score through a new guidance weight $\mu_t(\mathbf{c}_p, \mathbf{c}_S)$, which is used in $\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)$ defined in Eq. (3). This weight controls both the strength and the spatial positioning of the guidance at each timestep:

$$\mu_t(\mathbf{c}_p, \mathbf{c}_S) = \lambda(t) \cdot P_+(\mathbf{c}_p, \mathbf{c_S}) \cdot M(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) \tag{7}$$

where $P_+(\mathbf{c}_p, \mathbf{c_S}) = \max(0, P(\mathbf{c}_p, \mathbf{c_S}))$ to ensure only non-negative contributions influence the guidance. $\lambda(t)$ is a pre-defined function of timestep t, detailed later in this section. $M(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)$ acts as a mask that selectively applies guidance to regions likely to contain unsafe content, thereby promoting safe image generation. To elaborate on $M(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)$, we scale each mask value based on the pixel-wise proxy value, similar to the $\text{Sim}(\cdot,\cdot)$ function used in Eq. (6). Each pixel value for $M(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)$ is defined as follows:

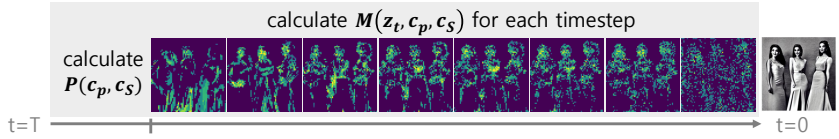


Figure 3: Guiding process of SP-Guard.

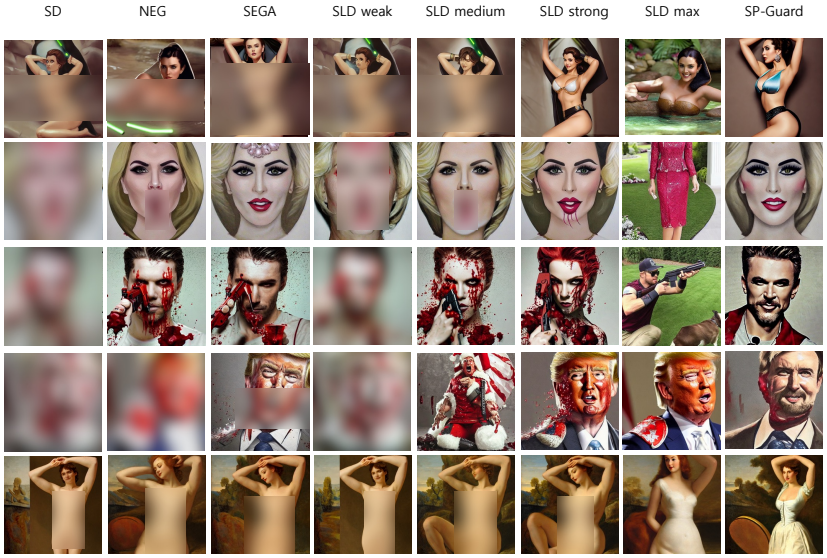


Figure 4: Qualitative comparison of methods for removing inappropriate content.

$$M(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)[i,j,k] = \begin{cases} 1 + \max(0, |\psi|) & \text{if } |\Delta\mathbf{c}_{S,t}[i,j,k]| > \eta_q(|\Delta\mathbf{c}_{S,t}|) \\ 0 & \text{else} \end{cases} \tag{8}$$

$$\text{with} \quad \psi = \text{Sim}(\Delta\mathbf{c}_{p,t}[i,j,:], \Delta\mathbf{c}_{S,t}[i,j,:]), \quad \text{where} \quad \eta_q(|\Delta\mathbf{c}_{S,t}|) = q\text{-percentile of } |\Delta\mathbf{c}_{S,t}|.$$

The masking condition is motivated by Brack *et al.* [16], who showed that the noise space consists of semantic concepts, with each concept concentrated in the upper and lower tails of the noise distribution. Accordingly, we mask the top q-percentile elements and compute cosine similarity for the corresponding pixels. In Fig. 3, we visualize $M(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)$, showcasing how our mask design strategically applies safe guidance to specific regions, such as nude body parts. In contrast, the masking process of SLD in Fig. 2b spreads the guidance over unrelated areas of the image, often altering benign content unnecessarily. This difference stems from the novelty of SP-Guard, which combines the prompt-adaptive risk score in Eq. (6) with the selective masking in Eq. (8), enabling more precise and targeted guidance.

Lastly, we use a step function as a default form of $\lambda(t)$ in Eq. (7). Specifically, after t_p timesteps, $\lambda(t)$ is set to λ_{max} and subsequently reduced to 1.0 in the later steps. The reduction is essential to avoid visual artifacts, as prior work [2] shows that no such artifacts or distortion occur when the guidance scale is capped at 1.0. Furthermore, Yi *et al.* [20] and Balaji *et al.* [21] observe that text prompts primarily influence the early diffusion steps, while the later stages focus on denoising and completing details using the latent image itself. These observations support our design: safe guidance should be prominent in the early steps, but does not require strong influence later on, and should be limited to preserve image quality. We validate the effectiveness of the step function in Section 3 and also explore the impact of varying λ_{max} or using alternative scheduling strategies for $\lambda(t)$.

3. Experiments

3.1. Experimental Setup

We compare SP-Guard with the original Stable Diffusion (SD) [1] and inference-time guiding methods: SD with a simple negative prompt (NEG), SEGA [16], and four configurations of SLD [10]. Since many existing methods [7, 22, 23, 24, 25, 26, 27, 28] report their results only on a single unsafe concept or handle different unsafe categories separately, their practical applicability is somewhat limited in multi-

Table 1
Results of safety, content preservation, and image quality. Each highlighted color corresponds to the best and second-best unsafe rates, as well as values worse than SP-Guard in terms of LPIPS, FID, or CLIP. SP-Guard achieves a strong trade-off, improving safety while preserving content and fidelity.

	I2P		Ring-A-Bell		MMA-Diffusion		UnlearnDiff		COCO-30k		DrawBench
	Unsafe ↓	LPIPS ↓	Unsafe ↓	LPIPS ↓	Unsafe ↓	LPIPS ↓	Unsafe ↓	LPIPS ↓	FID ↓	CLIP ↑	CLIP ↑
SD	25.16	–	78.51	–	67.25	–	27.20	–	19.36	0.310	0.308
NEG	14.68	0.45	72.70	0.46	57.45	0.44	19.77	0.43	24.69	0.301	0.298
SEGA	13.94	0.32	61.30	0.40	58.10	0.29	18.15	0.33	23.40	0.301	0.299
SLD-weak	19.45	0.17	74.69	0.19	63.98	0.14	23.41	0.18	20.65	0.308	0.305
SLD-medium	15.27	0.34	68.00	0.36	60.47	0.30	18.82	0.34	22.51	0.303	0.301
SLD-strong	11.56	0.45	52.73	0.47	52.18	0.43	15.51	0.44	25.83	0.296	0.292
SLD-max	9.97	0.56	33.83	0.56	41.92	0.54	13.36	0.54	33.85	0.288	0.282
SP-Guard	11.23	0.39	25.45	0.51	48.58	0.38	15.09	0.40	20.78	0.304	0.299

concept scenarios. Therefore, we primarily evaluate SP-Guard against SLD variants, as both address *multiple* unsafe concepts concurrently through a unified guidance process, enabling a fair comparison. We evaluate safe image generation on four datasets: I2P [10], Ring-A-Bell [5], MMA-Diffusion [6], and UnlearnDiff [29]. To assess image quality, we use DrawBench [2] and COCO-30k [30], which contain benign prompts. To assess the safety of generated images, we primarily report the unsafe content detection rate and its relative improvement over SD. We use an average score across four safety classifiers, MHSC [31], Q16 [32], NudeNet [33], and SD's built-in Safety-Checker [1], to provide a balanced estimate of overall harmfulness. To assess content preservation, we use LPIPS [34], which measures perceptual similarity between images generated by SD and each method, thereby quantifying how well the non-unsafe regions are retained. We also report CLIP-score [35] to evaluate image-text alignment and FID [36] to assess image fidelity. We use SD v1.4 with 50 diffusion steps and default settings across all baselines. For our method, $\lambda_{\max}$=4.0, q=0.9, and t_p=10, unless specified otherwise.

3.2. Qualitative analysis

The effectiveness of SP-Guard is demonstrated in Fig. 4. SP-Guard consistently generates safe images where SD fails. For example, it adds clothing in prompts involving nudity and replaces excessive blood in violent prompts with benign red elements. Unlike SLD, which applies guidance inconsistently, SP-Guard achieves reliable and prompt-adaptive safety through proxy-based guidance. Moreover, SP-Guard effectively confines guidance to areas identified as unsafe.

3.3. Quantitative results & Analysis

Evaluation results of safe image generation and content preservation across all datasets are shown in Table 1. As shown in the table, SP-Guard achieves safety performance comparable to SLD-max, ranking among the top inference-time guiding methods. However, it significantly outperforms SLD-max in image preservation. Notably, the LPIPS values of SP-Guard are comparable to baselines that exhibit minimal safety gains, highlighting the effectiveness of our selective masking strategy.

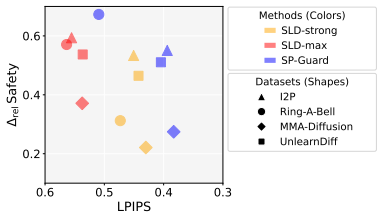


Figure 5: **Trade-off between safety improvement and content preservation.** Points further to the upper right indicate safer image generation with better content preservation.

We further evaluate the image quality using FID and CLIP scores on COCO-30k and DrawBench. As shown in the right-most columns of Table 1, SP-Guard achieves FID and CLIP scores closer to those of the original SD, while maintaining superior or comparable safe generation performance to SLD-max. Notably, SP-Guard outperforms the other baselines, except SLD-weak, which shows considerably lower performance in safety. To illustrate the trade-off between safety and content preservation, Fig. 5 shows the results for the top-performing methods: SLD-strong, SLD-max, and SP-Guard. The y-axis represents the average relative improvement over SD in unsafe detection rates, while the reversed x-axis shows the LPIPS, indicating perceptual similarity to images generated by SD. Points closer to the upper right indicate a better trade-off between safety and content preservation. SP-Guard consistently achieves

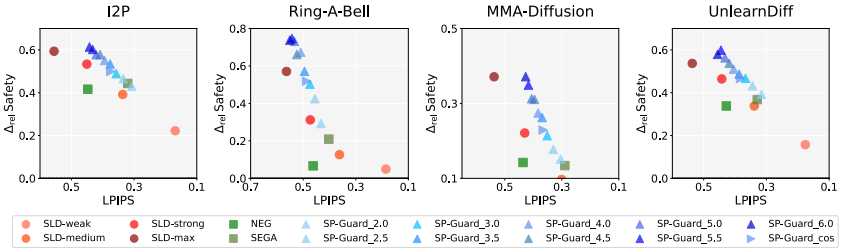


Figure 6: Evaluation of $\lambda(t)$ variations. This plot demonstrates the impact of varying λ of SP-Guard.

lower LPIPS scores than SLD-max and SLD-strong (except against SLD-strong on Ring-A-Bell), showing that the generated images by SP-Guard remain closer to the original SD outputs while ensuring safety.

We vary the maximum guidance scale λ_{max} from 2.0 to 6.0 in increments of 0.5, and also evaluate a cosine-based schedule as an alternative to the step function. Fig. 6 shows the results in the same format as the trade-off plot. SP-Guard consistently aligns with the Pareto front, showing robust performance across different λ_{max} values and scheduling strategies.

4. Discussion & Conclusion

This work highlights the importance of accurately estimating the potential harmfulness of generated content. Moreover, as SP-Guard is an inference-time guiding approach, it allows flexible modification or addition of unsafe concepts without retraining. Such adaptability enables rapid alignment with evolving social norms and regulations [37], making the method practical for real-world moderation pipelines and dynamic regulatory environments. Moreover, since our method relies on the general mechanism of guidance and the similarity between the intended semantics and harmful concepts, the framework can be naturally extended to other modalities such as video or speech generation. Beyond improving safety, our work strengthens the trustworthiness of generative AI systems in two ways. First, by diagnosing the failure modes of prior approaches, we emphasize the importance of carefully designing both the guidance mechanism and the masking process. Second, by estimating the prompt-specific potential harmfulness, SP-Guard offers transparency and controllability: users and deployers can see when and why safety interventions are applied. These features enhance trustworthiness rather than merely increasing safety. However, operating at inference time introduces some slowdown compared to standard SD. This could be mitigated by integrating recent advances in accelerating diffusion models [38, 39]. Finally, although SP-Guard reduces hyperparameter complexity compared to SLD, it still requires tuning values such as $\lambda(t)$ and q. A promising future direction is to dynamically adjust $\lambda(t)$ based on the guidance signal at each timestep. Despite these limitations, SP-Guard provides a lightweight, adaptable, and selective inference-time approach for safer text-to-image generation. Experiments on four unsafe-related datasets demonstrate significant improvements in safe generation with strong content preservation, while results on two benign datasets confirm its ability to maintain high fidelity. Looking ahead, we believe SP-Guard can be further enhanced and integrated with advancements in diffusion models, paving the way toward safe, responsible, and trustworthy AI.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant [No.2021R1A2C2007884] and by Institute of Information & communications Technology Planning & Evaluation (IITP) grants [RS-2021-II211343, RS-2021-II212068, RS-2022-II220113, RS-2022-II220959] funded by the Korean government (MSIT).

Declaration on Generative AI

During the preparation of this work, the author used Grammarly in order to check grammar and spelling. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Advances in neural information processing systems 35 (2022) 36479–36494.

[3] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, Advances in Neural Information Processing Systems 35 (2022) 25278–25294.

[4] A. Birhane, V. U. Prabhu, E. Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes, arXiv preprint arXiv:2110.01963 (2021).

[5] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, C.-Y. Huang, Ring-a-bell! how reliable are concept removal methods for diffusion models?, arXiv preprint arXiv:2310.10012 (2023).

[6] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, Q. Xu, Mma-diffusion: Multimodal attack on diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7737–7746.

[7] Y. Yang, R. Gao, X. Yang, J. Zhong, Q. Xu, Guardt2i: Defending text-to-image models from adversarial prompts, arXiv preprint arXiv:2403.01446 (2024).

[8] R. Liu, A. Khakzar, J. Gu, Q. Chen, P. Torr, F. Pizzati, Latent guard: a safety framework for text-to-image generation, arXiv preprint arXiv:2404.08031 (2024).

[9] Y. Qu, X. Shen, Y. Wu, M. Backes, S. Zannettou, Y. Zhang, Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images, arXiv preprint arXiv:2405.03486 (2024).

[10] P. Schramowski, M. Brack, B. Deiseroth, K. Kersting, Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22522–22531.

[11] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, D. Bau, Erasing concepts from diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2426–2436.

[12] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, D. Bau, Unified concept editing in diffusion models, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5111–5120.

[13] Z. Wu, H. Gao, Y. Wang, X. Zhang, S. Wang, Universal prompt optimizer for safe text-to-image generation, arXiv preprint arXiv:2402.10882 (2024).

[14] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, Advances in neural information processing systems 32 (2019).

[15] J. Ho, T. Salimans, Classifier-free diffusion guidance, arXiv preprint arXiv:2207.12598 (2022).

[16] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, K. Kersting, Sega: Instructing text-to-image models using semantic guidance, Advances in Neural Information Processing Systems 36 (2023) 25365–25389.

[17] R. Chavhan, D. Li, T. Hospedales, Conceptprune: Concept editing in diffusion models via skilled neuron pruning, arXiv preprint arXiv:2405.19237 (2024).

[18] Y. Dalva, P. Yanardag, Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24209–24218.

[19] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, A. Passos, Ledits++: Limitless image editing using text-to-image models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8861–8870.

[20] M. Yi, A. Li, Y. Xin, Z. Li, Towards understanding the working mechanism of text-to-image diffusion model, arXiv preprint arXiv:2405.15330 (2024).

[21] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, et al., ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, arXiv preprint arXiv:2211.01324 (2022).

[22] A. Heng, H. Soh, Selective amnesia: A continual learning approach to forgetting in deep generative models, Advances in Neural Information Processing Systems 36 (2023) 17170–17194.

[23] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, J. Lee, Safeguard text-to-image diffusion models with human feedback inversion, arXiv preprint arXiv:2407.21032 (2024).

[24] S. Lu, Z. Wang, L. Li, Y. Liu, A. W.-K. Kong, Mace: Mass concept erasure in diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6430–6440.

[25] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, W. Xu, Safegen: Mitigating unsafe content generation in text-to-image models, arXiv e-prints (2024) arXiv–2404.

[26] D. Chen, Z. Li, M. Fan, C. Chen, W. Zhou, Y. Li, Eiup: A training-free approach to erase non-compliant concepts conditioned on implicit unsafe prompts, arXiv preprint arXiv:2408.01014 (2024).

[27] C. Gong, K. Chen, Z. Wei, J. Chen, Y.-G. Jiang, Reliable and efficient concept erasure of text-to-image diffusion models, in: European Conference on Computer Vision, Springer, 2024, pp. 73–88.

[28] J. Yoon, S. Yu, V. Patil, H. Yao, M. Bansal, Safree: Training-free and adaptive guard for safe text-to-image and video generation, arXiv preprint arXiv:2410.12761 (2024).

[29] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, S. Liu, To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now, in: European Conference on Computer Vision, 2024.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[31] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, 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, 2023, pp. 3403–3417.

[32] P. Schramowski, C. Tauchmann, K. Kersting, Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content?, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1350–1361.

[33] P. Bedapudi, Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[36] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[37] I. Solaiman, Z. Talat, W. Agnew, L. Ahmad, D. Baker, S. L. Blodgett, C. Chen, H. Daumé III, J. Dodge, I. Duan, et al., Evaluating the social impact of generative ai systems in systems and society, arXiv preprint arXiv:2306.05949 (2023).

[38] A. Habibian, A. Ghodrati, N. Fathima, G. Sautiere, R. Garrepalli, F. Porikli, J. Petersen, Clockwork diffusion: Efficient generation with model-step distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8352–8361.

[39] Y.-H. Chen, R. Sarokin, J. Lee, J. Tang, C.-L. Chang, A. Kulik, M. Grundmann, Speed is all you need: On-device acceleration of large diffusion models via gpu-aware optimizations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4651–4655.