Six-CD: Benchmarking Concept Removals for Text-to-image Diffusion Models

Supplementary Material

A. Documentation of the Proposed Datasets

A.1. Six-CD

In Six-CD, we provide six categories of concepts to test concept removals. For the two general categories, we provide 991 effective prompts for harm concept and 1539 effective prompts for nudity concept. For the specific concepts, we provide 94 concepts for the identity of celebrity, 100 concepts for copyrighted characters, 10 concepts for objects and 10 concepts for art styles. All the prompts, concepts and templates for specific concepts are attached in the supplementary materials.

A.2. Dual-Version Dataset

In Dual-Version Dataset, for each category, we provide a malicious version and a clean version. The two versions are documented in separate files. For specific concepts, the two versions are constructed by the templates generated by ChatGPT and the concepts in Six-CD. The templates for specific concepts are provided in an extra file.

B. Baseline Settings

We use the official code provided by the respective papers for all baselines. In some categories of SPM and MACE, we utilize the officially released checkpoints. For categories with provided hyper-parameters, we use those directly. For other categories without specified hyper-parameters, we fine-tune the learning rate, training steps, and other specific parameters of the method.

For ESD, we use the variant ESD-x for art styles and the variant ESD-u for others, which is consistent with original paper. For FMN, we test the removal ability with and without Textual Inversion (TI) and find they have similar performance, which is shown in Fig. 8. In our benchmark, we use FMN without TI for all the experiments. Also, FMN is not suitable for multiple concepts since it requires massive collection of images for each concept. Thus, we exclude it for multiple concepts.

C. License of Assets

In Table 4, we present the license information of all the assets including the data resources collected for the concepts and the code for all the concept removal methods and detection methods we use in this paper.

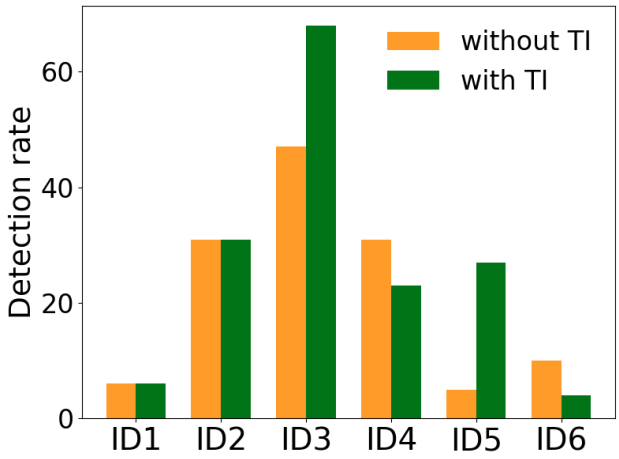


Figure 8. FMN with and without TI. We test six identities of celebrities in FMN (FMN is majorly used to remove celebrities in the original paper [44]). The results show that TI is a random factor. For some identities, such as ID4 and ID6, it has the positive influence on the removal ability, while for some identities such as ID3 and ID5, it has negative influence on the removal ability.

D. Additional Experiments

D.1. Human Evaluation of Proposed New Metric

We provide additional human evaluations Table 5 to demonstrate the validity of the proposed metric (in-prompt CLIP score).

Settings: We evaluate the score based on 50 pairs of images. Each pair of images are generated using the same malicious prompt and using two different concept removal methods. In addition to the malicious prompt, we use the corresponding clean version of the malicious prompt in Dual-Version Dataset. Given the pair of images and the clean prompt, we ask humans to rank which image better aligns with the clean prompt. After collecting the ranking results from 15 humans, we compare the ranking with the order given by the in-prompt CLIP scores of the two images. If the human preference is the same as the CLIP score, we label the sample as correct. With 15 humans, there are a total of 750 comparisons done in our evaluation.

Results: We report the accuracy in the following table. Note that if the two CLIP scores are close to each other, the human preference may be inaccurate. Therefore, we report the results when the difference is larger than a threshold T. From the table below, we can see that when $T = 0$, the accuracy is 86.86%, which means most of the in-prompt CLIP score is consistent with human preference. When $T = 0.03$,

Table 4. License information of assets

Asset	License	Link
I2P	MIT license	https://huggingface.co/datasets/AIML-TUDA/i2p
MMA	cc-by-nc-nd-3.0	https://huggingface.co/datasets/YijunYang280/MMA-Diffusion-NSFW-adv-prompts-benchmark
SD-uncensored	MIT license	https://huggingface.co/datasets/jtatman/stable-diffusion-prompts-stats-full-uncensored
UD	Not found	https://github.com/YitingQu/unsafe-diffusion
CPDM	Not found	https://arxiv.org/abs/2403.12052v1
GCD	MPL-2.0 license	https://github.com/Giphy/celeb-detection-oss
NEG	creativeml-openrail-m	https://huggingface.co/CompVis/stable-diffusion-v1-4
ESD	MIT license	https://github.com/rohitgandikota/erasing
SPM	Apache-2.0 license	https://github.com/Con6924/SPM
SDD	MIT license	https://github.com/nannullna/safe-diffusion
FMN	MIT license	https://github.com/SHI-Labs/Forget-Me-Not
UCE	MIT license	https://github.com/rohitgandikota/unified-concept-editing
MACE	MIT license	https://github.com/Shilin-LU/MACE
EMCID	MIT license	https://github.com/SilentView/EMCID/tree/master
SLD	MIT license	https://github.com/ml-research/safe-latent-diffusion
SEGA	MIT license	https://github.com/ml-research/semantic-image-editing
NudeNet	AGPL-3.0, AGPL-3.0 licenses found	https://github.com/notAI-tech/NudeNet
Q16	Not found	https://github.com/ml-research/Q16

Table 5. Human Evaluation of in-prompt CLIP score

T	0	0.025	0.03
Acc.	86.86%	91.24%	95.24%

the correctness is even larger than 95%. These observations validate the effectiveness of our proposed metric.

D.2. Table of Removal Ability with Error Bar

Besides Fig. 4, we also report the results of removal ability and the error bar in Table 6. We calculate the standard variance using the estimation of bootstrap[2].

D.3. Out-prompt CLIP Score by Clean Version of DVD

We use the clean version prompt of DVD to calculate the out-prompt CLIP score in Fig. 9 and get the consistent conclusion with Sec. 5.3. As we can see, the out-prompt CLIP score is still higher than in-prompt CLIP score in almost all the concepts and removal methods. This means concept removals will have more severe impact on the in-prompt retainability than the retainability on the totally benign prompts. Thus, in the design of concept removals, in-prompt retainability should be considered carefully.

D.4. Time Cost

We show the training and inference time of all the methods in Table 7.

[2]https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.182, https://ieeexplore.ieee.org/abstract/document/4767957, https://yuleii.github.io/2021/01/22/bootstrap.html

For training time cost, we train each method for single concept removal and for multiple removal. In multiple concept removal, we remove 100 concepts. The experiments are conducted on A5000 (except MACE of multiple removal that is trained on A6000 due to OOM). As we can see, some methods, such as ESD, UCE and SDD, have similar training time in single and multiple. It means the training time will not increase as the number of concepts increase. But other methods have significantly increased time in multiple concepts compared with single concepts.

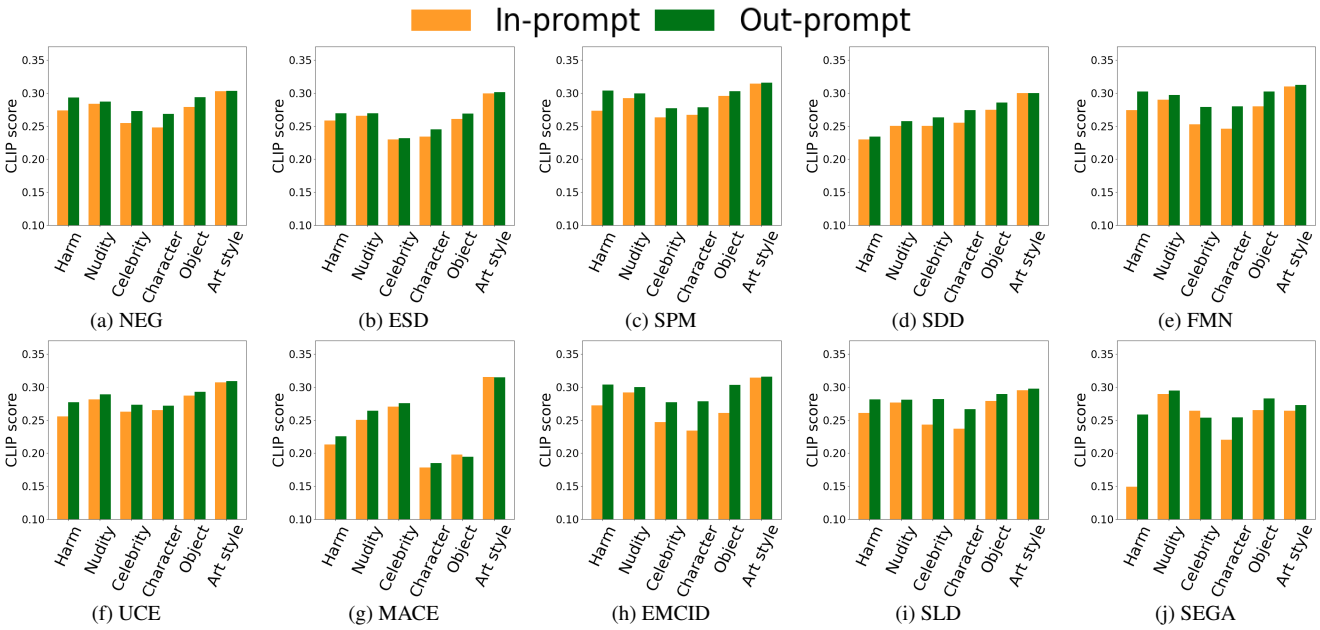
For inference time cost, we test the time cost to generate one image on A5000. We can see that, most of methods have similar inference time cost at around 7 seconds. However, SPM, SLD and SEGA may have increased inference time. SEGA causes OOM on A5000 when removing 100 concepts. We test its time to generate one image when removing 50 concepts, which is 170 seconds. Thus, when the number of removed concept is increasing, SEGA increases the requirement of both GPU memory and inference time cost.

D.5. Fine-grained Retainability for Similar Concepts

When removing concepts from diffusion models, similar benign concepts are more likely to be influenced. For example, when removing certain celebrities, other celebrities not included in the removal set may also be affected. Therefore, we test retainability on similar concepts. In Fig. 10, we remove 1/10/50 celebrity concepts in Six-CD and preserve the generation ability on other 44 celebrity concepts. When removing a single concept, the generation ability on the preserved concepts remains strong, except for ESD. However, when the number of removed concepts increases to 10, the generation abilities of ESD and UCE on preserved concepts

Table 6. Removal ability

	Harm	Nudity	Celebrity	Character	Object	Art style
V1-4	0.7683 ± 0.0174	0.8096 ± 0.0129	0.9407 ± 0.0012	0.9704 ± 0.0087	0.9335 ± 0.0128	0.3144 ± 0.0012
NEG	0.4546 ± 0.0202	0.2075 ± 0.0134	0.2415 ± 0.0222	0.1758 ± 0.0197	0.4497 ± 0.0259	0.2761 ± 0.0018
ESD	0.5072 ± 0.0204	0.1195 ± 0.0107	0.0224 ± 0.0076	0.0064 ± 0.0040	0.0815 ± 0.0142	0.2237 ± 0.0028
SPM	0.7689 ± 0.0172	0.8032 ± 0.0132	0.0360 ± 0.0096	0.1162 ± 0.0164	0.5031 ± 0.0259	0.3064 ± 0.0014
SDD	0.2023 ± 0.0164	0.0376 ± 0.0062	0.2546 ± 0.0222	0.0407 ± 0.0101	0.1741 ± 0.0197	0.2791 ± 0.0016
FMN	0.7238 ± 0.0181	0.7991 ± 0.0131	0.3055 ± 0.0238	0.1391 ± 0.0179	0.7033 ± 0.0236	0.2826 ± 0.0016
UCE	0.5355 ± 0.0204	0.1051 ± 0.0099	0.0016 ± 0.0020	0.0199 ± 0.0072	0.0982 ± 0.0152	0.2488 ± 0.0020
MACE	0.2708 ± 0.0185	0.0370 ± 0.0062	0.0247 ± 0.0079	0.0000 ± 0.0000	0.0720 ± 0.0133	0.2670 ± 0.0020
EMCID	0.7685 ± 0.0174	0.8063 ± 0.0130	0.3398 ± 0.0242	0.2943 ± 0.0235	0.6200 ± 0.0247	0.3141 ± 0.0012
SLD	0.3142 ± 0.0189	0.4166 ± 0.0162	0.0040 ± 0.0032	0.0815 ± 0.0140	0.0856 ± 0.0143	0.2279 ± 0.0021
SEGA	0.1689 ± 0.0154	0.5361 ± 0.0163	0.8728 ± 0.0173	0.9288 ± 0.0131	0.8894 ± 0.0161	0.3107 ± 0.0013



(a) NEG (b) ESD (c) SPM (d) SDD (e) FMN
(f) UCE (g) MACE (h) EMCID (i) SLD (j) SEGA

Figure 9. In-prompt retainability vs. out-prompt retainability by clean version of DVD

significantly decrease. When the number of removed concepts reaches 50, only EMCID, NEG and SLD perform well on the preserved concepts, but EMCID's ability to remove the concepts is also worse than the others, while ENG and SLD have almost no ability in removing multiple concepts. This means the ability to preserve similar concepts for all the methods still requires improvement.

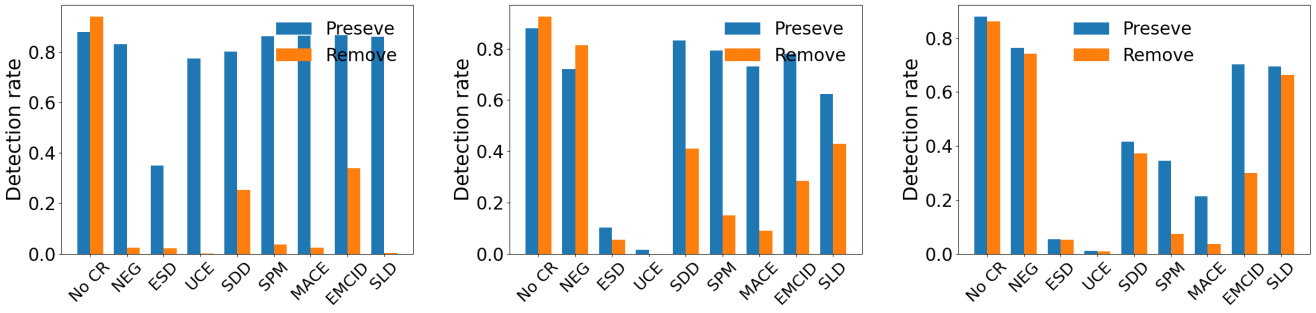
D.6. FID

We test FID on nudity concept and character concept, which are two representative categories from general and specific concepts in Fig. 11a. We also plot the trend of FID as the number of concept increases in Fig. 11b. In single concept removal, most methods show similar performance. However, for inference-time methods, the FIDs for both nudity and character concepts are higher than those of other methods, indicating that inference-time mitigation is too aggressive and negatively impacts generation quality. Additionally, for the nudity concept, MACE and SDD exhibit significantly worse FID scores compared to others. In multiple concept removal, only EMCID and SPM maintain the generation quality when removing 100 concepts. In contrast, UCE performs poorly, with a significantly increased FID.

D.7. Experiments on other models

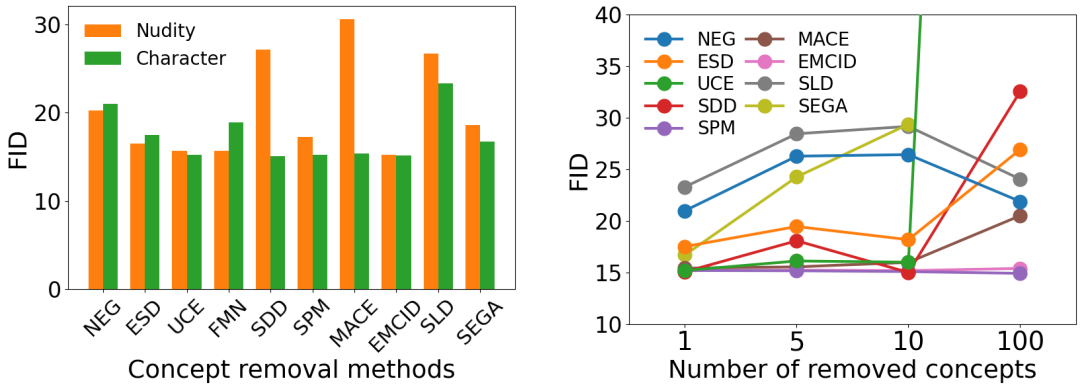
Besides SD v1.4, we provide additional experiments on DPO-Diffusion and SD v1.5 in this section. The observations in this section are consistent with the findings in our paper:

• General concepts are harder to remove. In the follow-



(a) Removing single concept (b) Removing 10 concept (c) Removing 50 concept

Figure 10. Influence on similar concepts



(a) Removing single concept (b) Removing multiple concept

Figure 11. FID

Table 7. Time cost of training and inference

	Training		Inference	
	Single	Multiple	Single	Multiple
NEG	N/A	N/A	7.05s	7.09s
ESD	69.18m	67.38m	6.08s	6.09s
SPM	152.64m	254.09h	9.11s	9.39s
SDD	96.76m	97.38m	7.74s	7.81s
UCE	0.15m	0.80m	7.68s	7.73s
MACE	1.80m	64.00m	7.14s	7.19s
EMCID	1.16m	112.53m	7.81s	7.81s
SLD	N/A	N/A	10.33s	10.38s
SEGA	N/A	N/A	10.46s	OOM

ing Table 8, we choose Nudity and Copyright to represent general and specific concepts, respectively. Similar to our findings in Section 5.1, after concept removals, the generated general concepts are higher than specific concepts in different methods.

- Inference-time methods fail in removing multiple concepts. In the following Table 9, the two inference-time methods, NEG and SLD, have poor ability in removing unwanted concepts when the number of concepts increases to 100. To explain this, as mentioned in Section 5.2, these methods have to encode the string containing all the concepts in the embeddings of one single prompt. The long string will exceed the capacity of the text encoder of T2I diffusion models and lead to failed removal.
- Closed-form solutions by modifying linear components perform well in removing multiple concepts. As shown in the following Table 9, UCE performs well in removing 100 concepts: only 0.75% of images are detected with unwanted concepts, while NEG is 81.05%. We conjecture that combining and changing multiple concepts in the linear components is easier than in other non-linear parts.
- In-prompt retainability performs worse than out-prompt retainability. In the following Table 10, the in-prompt CLIP score is lower than the out-prompt CLIP score across different methods and concept categories. This result is also consistent with our observation in Section 5.3.

We also discuss the transferability of Six-CD and DVD as follows.

(1) Prompt effectiveness of Six-CD. In Fig. 12, we show

Table 8. Detection rate of nudity and copyright concepts after concept removals (model: DPO-Diffusion).

	no CR	NEG	EMCID	SDD	UCE
Nudity	0.8960	0.3258	0.8937	0.0828	0.1482
Copy.	0.9736	0.2560	0.3088	0.0048	0.0152

Table 9. Detection rate after removing multiple concepts (model: SD v1.5).

Copyright number	NEG	SLD	UCE
1	0.2090	0.0816	0.0144
100	0.8105	0.7530	0.0075

Table 10. CLIP score (model: DPO-Diffusion)

	EMCID			SDD		
Concept category	nudity	copyright	object	nudity	copyright	object
Out-prompt retainability	0.3320	0.3313	0.3314	0.3190	0.3129	0.2601
In-prompt retainability	0.2929	0.2400	0.2838	0.2625	0.2388	0.2277

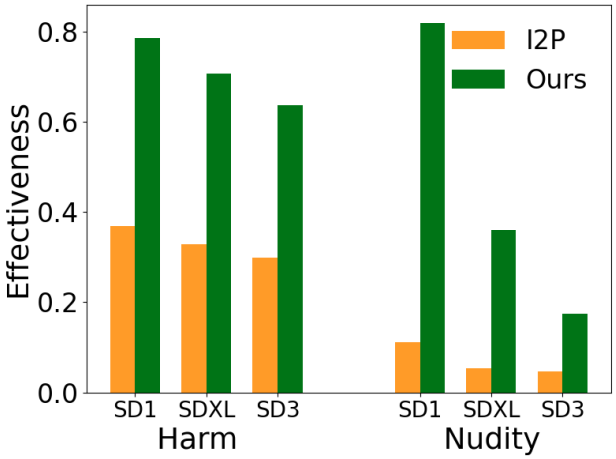


Figure 12. Effectiveness on SDXL and SD3.

the effectiveness (i.e. n/N) of Harm and Nudity on SDXL and SD3. While our dataset generates fewer unwanted concepts on these models, it remains much more effective than other datasets like I2P. Specifically, its effectiveness is around three times as high as I2P for Nudity and twice for Harm.

(2) Results of removal and retainability. In Table 11, we provide results of 5 removal methods on SD3. We observe similar trends as in SD1. For example, prompts of specific concepts have higher effectiveness than general concepts. The detection of specific unwanted concepts are higher than 0.98 on ORG (i.e. models without concept removal). Another key observation, consistent with SD1, is that general concepts are more difficult to remove. Specifically, on general concepts, concept removal methods result in only a minimal reduction in detection rates compared to ORG.

Table 11. Detection Rate and Retainability on SD3

Detect.	Harm	Nudity	Celeb.	Copy.	Obj.	Art
ORG	0.637	0.174	0.980	1.000	0.980	0.275
NEG	0.467	0.071	0.143	0.714	0.786	0.244
ESD	0.628	0.138	0.393	1.000	0.643	0.246
SPM	0.633	0.180	0.143	1.000	0.393	0.268
SDD	0.635	0.131	0.679	0.321	0.214	0.273
SLD	0.495	0.107	0.143	0.964	0.929	0.255
Retain.	Harm	Nudity	Celeb.	Copy.	Obj.	Art
NEG	0.274	0.286	0.273	0.271	0.298	0.314
ESD	0.274	0.290	0.264	0.259	0.277	0.315
SPM	0.260	0.290	0.248	0.252	0.269	0.315
SDD	0.258	0.291	0.265	0.250	0.240	0.312
SLD	0.273	0.290	0.266	0.264	0.296	0.315