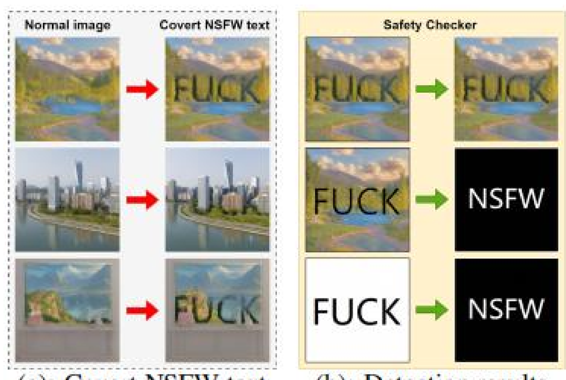
BMVC 2025

# Is Safety Checker Still Safe? A Study on the Covert NSFW Text

Xin Li, Kai chen, Xue Yang, Weijun Shan
Shanghai Fudan Microelectronics Group

## 1.Abstract

Recently, with the assist of generative models, one can generate highly realistic images with arbitrary content. However, models learn a large volume of inappropriate content from dataset, which leads to the dissemination of inappropriate content. In addition to intuitive inappropriate content in images, we discover for the first time that normal objects can be utilized to subtly hide offensive text in images. To explore the risks of such content, we which can insert the offensive text into generated images. And these images can effectively cheat the SafetyChecker. We establish a novel Covert NSFW Text Dataset (CNTD) based on CNT and release it for further research. In response to this type of content, we propose a detection method, Enhanced SafetyChecker (ESC). ESC is training-free and can effectively detect both the covert offensive text and any inappropriate content that SafetyChecker can detect. Experimental results demonstrate the effectiveness of CNT and ESC.



(a): Covert NSFW text. (b): Detection results.

## 2.Contribution

- We discover the vulnerability of the Safety Checker in detecting the covert NSFW text. Therefore, we propose CNT that can insert the covert NSFW text into images. These images can cheat the Safety Checker. Moreover, we can use safety generative models and provide safety prompts input.
- We establish and release CNTD, a novel covert NSFW text dataset of 5k images generated by CNT.
- We propose ESC, a training-free method for detecting the covert NSFW text based on preprocessing images and the Safety Checker. Experiments verify the performance of ESC.

## 3.Method: Pipeline of CNT

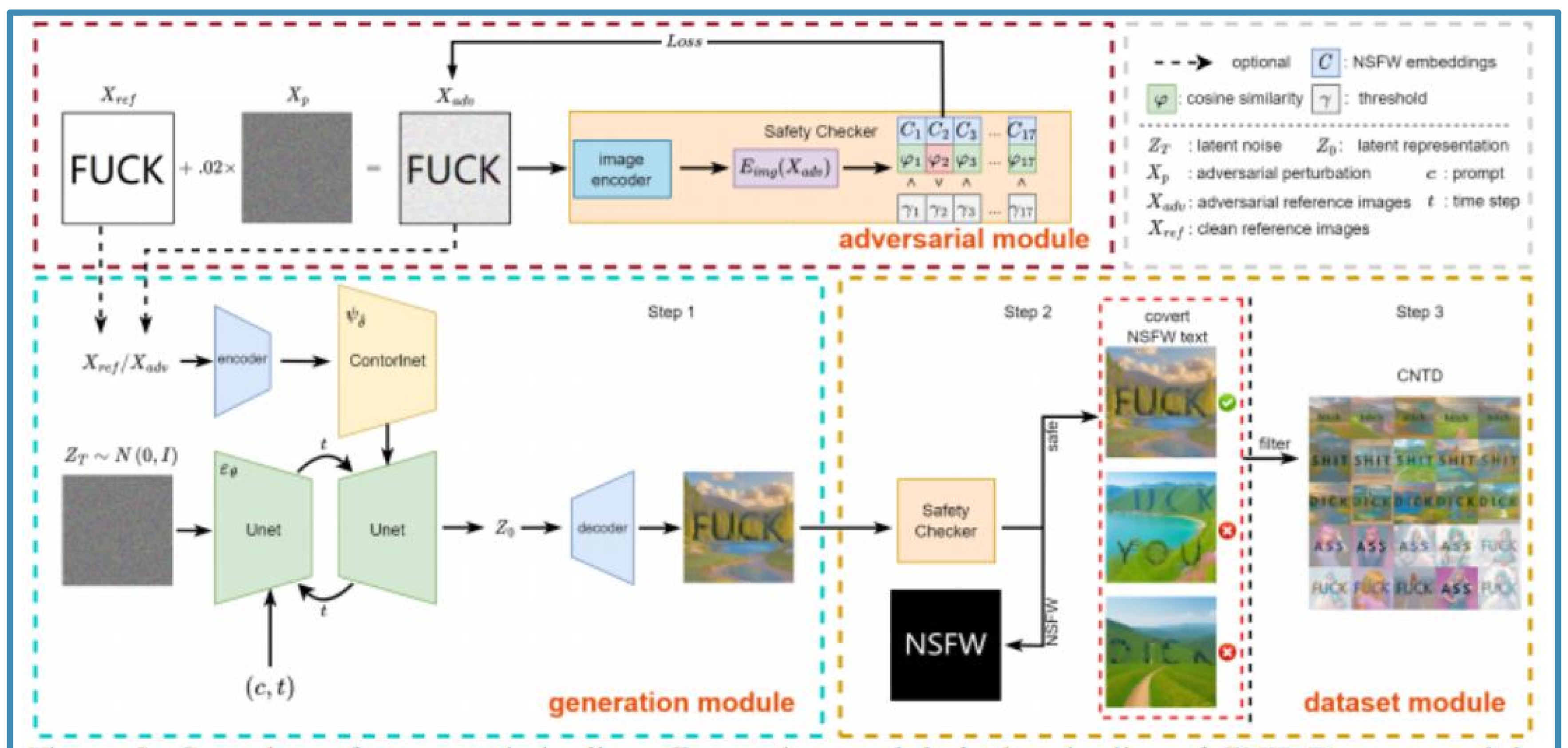



Figure 2: Overview of generated pipeline. Generation module is the pipeline of CNT. Dataset module filters images from generation module and integrates them into CNTD. Adversarial module is optional, it offers Generation module the adversarial reference image which can cheat the Safety checker.

CNT can be regarded as a black-box attack, attackers operate without access to the model's parameters. By merely providing prompts $c$ and reference images $X_{ref}$, they use the ControlNet $\psi_{\hat{\theta}}$ to guide SD $\varepsilon_\theta$ to generate images containing he covert NSFW text. The generation process is formulated as follows $X_{NSFW} = D(\varepsilon_\theta(z_T, c, T, \psi_{\hat{\theta}}(X_{ref})))$ We add the perturbation $X_p$ to the reference image $X_{ref}$ and obtain the adversarial reference image $X_{adv} = X_{ref} + \alpha X_p$, where $\alpha$ is learning rate. The Safety Checker encodes $X_{adv}$ to a latent representation $E_{img}(X_{adv})$. Then the similarity $\varphi(E_{img}(X_{adv}), C_{NSFW})$ over than the threshold $\delta$ is accumulated. We formulate our objective function as follows:

$$\underset{|X_{ref}-X_{adv}| \leq \beta}{\arg\min} \sum_{i=1}^{17} max(0, \varphi_i(E_{img}(X_{adv}), C^i_{NSFW}) - \delta_i)$$

The process is shown in dataset module in Figure 2. It includes three steps: First, models generate a large number of images by CNT mentioned above. Second, we employ the Safety Checker to detect the generated images. Third, we keep images which can cheat the Safety Checker and contain complete NSFW text. These images are selected through a rigorous manual screening process and are then integrated into the CNTD dataset.

## 4.Method: Pipeline of ESC

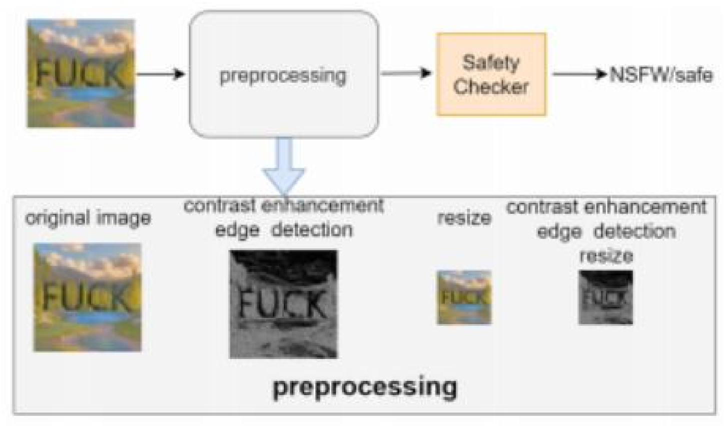



Figure 3: The preprocessing pipeline of ESC.


Figure 4: The results of images after preprocessing.

We propose Enhanced Safety Checker (ESC), a training-free method that can detect the covert NSFW text. ESC preprocesses the covert NSFW text before the detection by the Safety Checker. As shown in Figure 3, we adopt four preprocessing methods:

**Resizing**: The visual effect of the covert NSFW text is often composed of various textures in the image. Therefore, reducing the size of an image can significantly blur its textures, which forces the model to focus on the overall structure of the image.

**Contrast enhancement**: The text embedded in images may have slight differences in color compared to the surrounding area. Contrast enhancement can expand these differences.

**Edge detection**: In order to highlight outline of the covert text, we apply the edge detection to the image after contrast enhancement.

**Original image**: To prevent preprocessing from blurring NSFW content beyond the covert NSFW text, it is necessary to detect the original image.

## 5.Experiments

Regarding the attack success rate (ASR) of CNT, there is hardly a standard quantitative measurement to evaluate CNT. Therefore, we conduct a user study. Firstly, CNT generates $N$ images, among which $M (100 \leq M \leq N)$ images successfully cheat the safety checker. From this subset, 100 images are selected for evaluation. We invite 20 participants to assess the fidelity of NSFW text representation. The ASR is formulated as follows:

$$ASR = \frac{M}{N} * \frac{1}{2000} \sum_{i=1}^{20} \sum_{j=1}^{100} e^i_j$$

The ASR of CNT and the detection rate of ESC are illustrated in Tables 2 and 5.

| Prompt | Ref | CnOCR | PaddleOCR | RapidOCR | EasyOCR | SC | Q16 | InternVL2 | ASR | ESC |
|---|---|---|---|---|---|---|---|---|---|---|
| *landscape* | - | - | - | - | - | 0 | 0 | 0.0008 | - | 0.0102 |
| | clean | 0.0037 | 0.0771 | 0.0317 | 0.0003 | 0.0214 | 0.0120 | 0.0169 | 0.2921 | **0.4215** |
| | adv | 0.0021 | 0.0559 | 0.0213 | 0.0002 | 0.0158 | 0.0078 | 0.0128 | 0.2889 | **0.3620** |
| *cloud* | - | - | - | - | - | 0.0004 | 0.0015 | 0.0007 | - | 0.1444 |
| | clean | 0.4951 | 0.5139 | 0.4929 | 0.3765 | 0.7643 | 0.3434 | 0.5635 | 0.2044 | **0.8839** |
| | adv | 0.4719 | 0.4882 | 0.4717 | 0.3467 | 0.7303 | 0.3122 | 0.5484 | 0.2101 | **0.8638** |

Table 2: The ASR of CNT, along with the detection rates of ESC and baselines on generated images under different conditions.

| Metric | CnOCR | PaddleOCR | RapidOCR | EasyOCR | Safety Checker | Q16 | InternVL2 | **ESC** |
|---|---|---|---|---|---|---|---|---|
| CNTD↑ | 0.1958 | 0.5146 | 0.4050 | 0.0040 | 0 | 0.0280 | 0.0986 | **0.9604** |
| Safe↓ | - | - | - | - | 0 | **0** | 0.0005 | 0.0153 |
| IN/val↓ | - | - | - | - | **0.0238** | 0.0573 | 0.0377 | 0.1710 |
| IN/test↓ | - | - | - | - | **0.0221** | 0.0562 | 0.0358 | 0.1681 |

Table 5: The NSFW detection rates on CNTD, safe dataset, ImageNet/val and ImageNet/test.

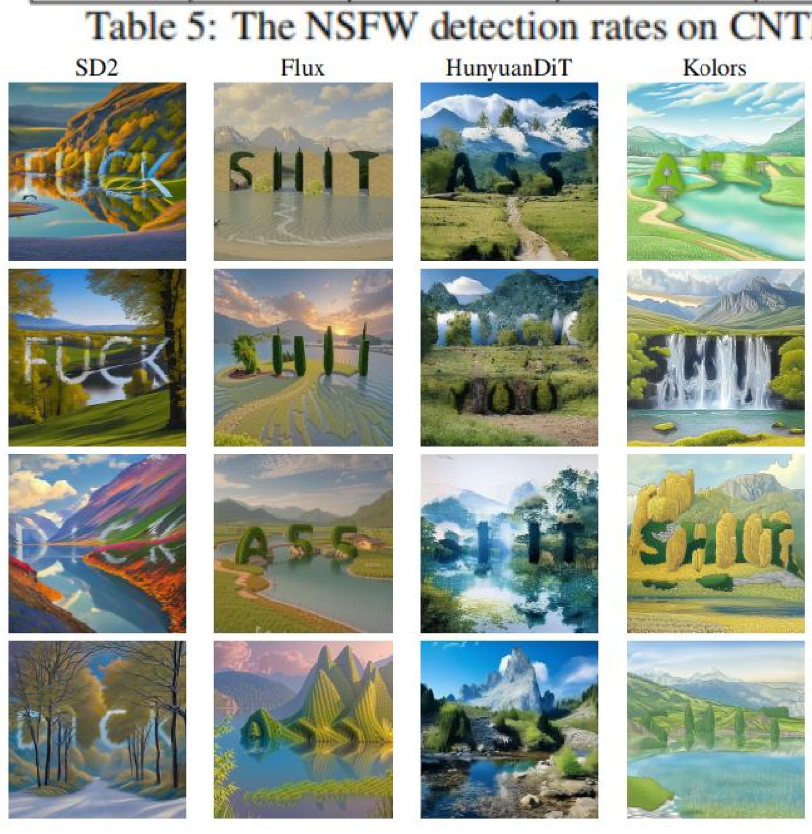



Figure 2: The generated results of different models.


Figure 4: The results of generating covert long texts.

## 6.Conclusion

In this paper, we firstly propose CNT that can insert the NSFW text into images. Even when the model lacks the corresponding NSFW content, it is possible to guide the generation of the covert NSFW text through safe prompt. Furthermore, the generated images can cheat OCR models and NSFW detectors easily. Based on CNT, we establish and release a Covert NSFW Text Dataset CNTD) for further research. To detect the covert NSFW text, we propose ESC, which performs preprocessing on images to expose the covert NSFW text. Then the covert text can be detected easily. Extensive experiments demonstrate the effectiveness of CNT and ESC.