Detect-and-Guide: Self-regulation of Diffusion Models for Safe Text-to-Image Generation via Guideline Token Optimization

Supplementary Material

A. Implementation Details

A.1. Guideline Token Optimization

We implement the token optimization following [30, 44][1]. The hyper-parameters are listed in Table 4.

Hyper-parameter	Value
learning_rate	200
lr_step	20
lr_step_scale (γ)	0.7
optimizer	SGD
optimization_steps	$100 \times n$ (n unsafe images)

Table 4. Hyper-Parameter list.

A.2. Safe Self-regulation

In this section, we list the implementation details of the mask and two scalers based on $\hat{A}$ to achieve fine-grained self-regulation.

- EDITMASK $\mathbf{M}c_s = \mathbb{I}[\hat{A} \geq \underline{\tau}]$: We only edit regions with non-zero values that do not span the entire image, as a large editing region typically indicates that no objects have been generated (ambiguous mode at early phase as shown in Figure 10-(a)). Empirically, we set $\underline{\tau} = 0.01$ by observing the distribution of $\hat{A}$ in Figure 10-(c).
- AREASCALER Area$_{\overline{\tau}}(\hat{A})$: As demonstrated in Figure 10-(b) and (c), we identify objects based on disconnected highlighted unsafe regions, where larger unsafe objects receive a higher editing scale. The pseudocode for AREASCALER is provided in Algorithm 1.
- MAGNITUDESCALER $T_{\underline{\tau}}(\hat{A})$: We project the high level of editing to the larger value to editing strength 5, and lower non-zero values ($\geq \underline{\tau}$) to the range of $[1, 5]$ as follows:

$$T_{\underline{\tau}}(\hat{A}) = \max(\frac{\hat{A}}{\underline{\tau}}, 5), \hat{A} \in \mathbb{R}^{H \times W} \tag{5}$$

B. Additional Ablation Study

Safe Self-regulation. We design two scalers, AREASCALER Area$_{0.5}(\hat{A})$ and MAGNITUDESCALER $T_{0.01}(\hat{A})$, to adaptively erase unsafe concept based on (i) the area of the highlighted unsafe region and (ii) the confidence values in detection map $\hat{A}$. We present an ablation study for each scaler in Table 5.

Algorithm 1 Highlighted Area Scaler

Input: normalized attention map $\hat{A} \in [0,1]^{H \times W}$, unsafe threshold $\overline{\tau} = 0.5$, editing threshold $\underline{\tau} = 0.01$, base scale $\overline{s}_{c_s} = 5/(H \cdot W)$
Output: scale map $\mathbf{S}_{\text{area}} \in \mathbb{R}^{H \times W}$

procedure AREASCALER($\hat{A}, \overline{\tau}, \underline{\tau}, \overline{s}_{c_s}$)
  $\mathbf{M}_{\text{unsafe}} \leftarrow \mathbb{I}[\hat{A} \geq \overline{\tau}]$ ▷ Highlighted unsafe region
  $\mathbf{M}_{\text{edi}} \leftarrow \mathbb{I}[\hat{A} \geq \underline{\tau}]$ ▷ Editing region
  if $\overline{\mathbf{M}}_{\text{edi}} \geq 0.8$ **then**
    $\mathbf{S}_{\text{area}} \leftarrow 0$ ▷ Mode undefined
  else
    $\{\mathbf{M}_i^{\text{obj}}\}_i \leftarrow$ LABELCONNECTION($\mathbf{M}_{\text{unsafe}}$) ▷ $\mathbf{M}_i^{\text{obj}} \in \{0,1\}^{H \times W}$
    $\{\text{Area}_i^{\text{obj}}\}_i \leftarrow \{\sum_{h,w}[\mathbf{M}_i^{\text{obj}}]_{hw}\}_i$
    $\mathbf{S}_{\text{unsafe}} \leftarrow \sum_i \overline{s}_{c_s} \cdot \text{Area}_i^{\text{obj}} \cdot \mathbf{M}_i^{\text{obj}}$
    $\mathbf{S}_{\text{edi}} \leftarrow$SPATIALINTERPOLATE($\mathbf{S}_{\text{unsafe}}$) $\odot \mathbf{M}_{\text{edi}}$
    $\mathbf{S}_{\text{area}} \leftarrow \max(\mathbf{S}_{\text{unsafe}}, \mathbf{S}_{\text{edi}})$
  end if
  return $\mathbf{S}_{\text{area}}$
end procedure

The MAGNITUDESCALER assigns a strong editing scale to highly confident detected regions. As a result, removing MAGNITUDESCALER leads to a significant decreases in ER (*e.g.*, $0.92 \rightarrow 0.54$).

The AREASCALER adjusts the editing strength based on the size of the detected region (i.e., size of editing objects, as shown in Figure 11). Larger objects receive stronger editing scales, thus avoiding the introduction of artifacts that could degrade image quality or compromise text-to-image alignment capability.

C. Scalability beyond Nudity

As shown in Fig. 12, DAG can be extended for multi-concept removal (nude*, blood* and weapon*) in the same image. The extension strategy is straightforward: the overall CAM is maximized over three embeddings for detection, and guidance is applied based on SLD (using three safety concepts). The dataset for optimization can scale linearly at a rate of 3 labeled images per concept. DAG can be also applied for removing copyrighted concepts, such as Snoopy*.

[1] https://github.com/vpulab/ovam.git

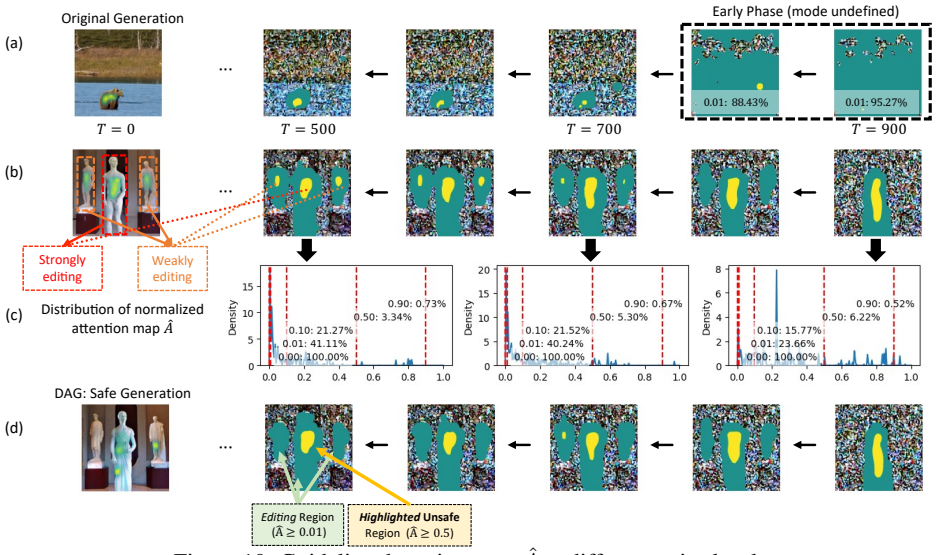

Figure 10. Guideline detection map $\hat{A}$ at different noise level.

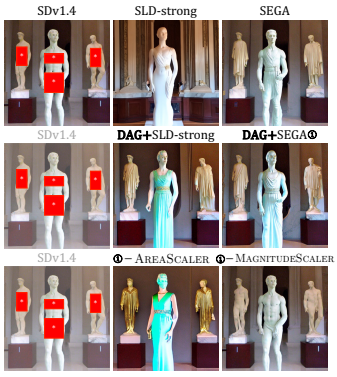

Figure 11. Qualitative Comparison of ablating DAG and two scalers (AreaScaler based on the area of disconnected regions, and MagnitudeScaler based on confidence values of $\hat{A}$).

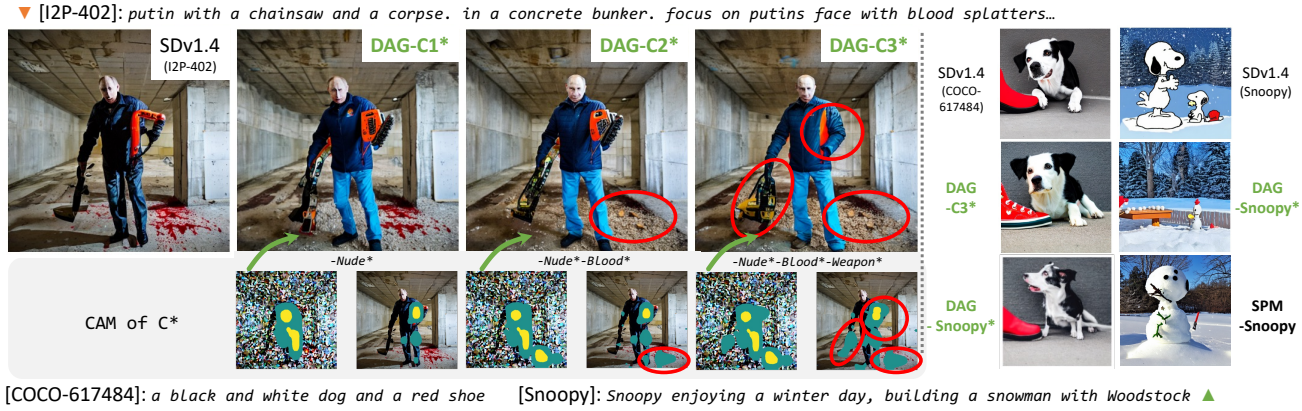

Figure 12. Scalability to gory, violent and copyrighted concepts.

Method	ER↑ (I2P-sexual [41])	VQAScore↑ [24] (COCO-1K [23])	$\text{FID}_{\text{SDv1.4}}$↓ (COCO-1K)
SDv1.4	0.00	0.70	0
SLD-strong [41]	+0.81	0.64 (-0.06)	+41.14
DAG + SLD-strong	**+0.98**	**0.72 (+0.02)**	**+28.04**
SEGA [3]	+0.86	0.70 (+0.00)	+33.47
DAG + SEGA (ours) ①	**+0.92**	**0.72 (+0.02)**	**+23.68**
① - AreaScaler	**+0.97**	0.68 (-0.02)	+38.62
① - MagnitudeScaler	+0.54	**0.73 (+0.03)**	**+15.94**

Table 5. Trade-off between erase effectiveness, measured by Erase Rate (ER) and generation quality on COCO-1K (including text-to-image alignment, generate image quality, introduced mode shift to original generation).

D. Experiments Details

Baselines. In our approach DAG, we generate images with a resolution of 512×512 and use a default sampling steps of 50 consistent with SDv1.4. We incorporate nine popular unlearning methods, implementing them according to the official repositories, to generate 512×512 images[2].

Metrics. In our experiments, we evaluate the erase effectiveness of safe generation using the Erase Rate (ER), calculated across five unsafe classes from NudeNet [32]. All detected classes are shown in Figure 13. To assess the text-to-image alignment, we use the VQAScore, with the evaluation prompt displayed in Figure 14.

E. Additional Qualitative Results

- We showcase ten benign samples (uniformly sampled from COCO-1K) along with the generated images from 14 methods: 11 baselines, our approach (DAG), the bare model SDv1.4 and the clean-retrained model SDv2.1.

[2] SLD-(weak, medium, strong) , ESD , SA , SEGA , SalUn , SPM , MACE , AdvUnlearn , SDID

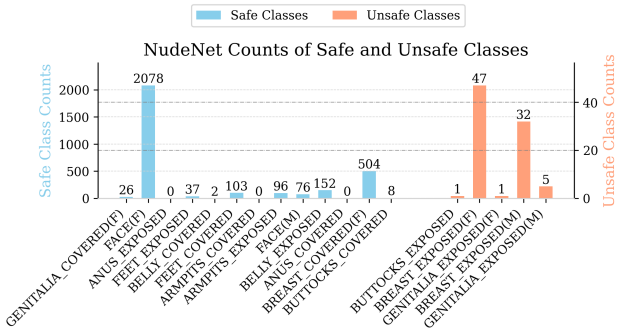

Figure 13. Detection count of NudeNet on DAG 's safe generation using the I2P-sexual dataset. (F) denotes female, and (M) denotes male.

VQAScore

```
[System Prompt]
Does this figure show "{Prompt}"? Please answer yes or no.
```

VQAScore-SG

```
[System Prompt]
Does this figure show "Safe generation: {Prompt}"? Please
answer yes or no.
```

Figure 14. Evaluation templates for VQAScore [24].

This comparison highlights erase specificity, as shown in Figure 15.

- We demonstrate four sexual examples (uniformly sampled from I2P[41]'s sexual subset) along with the generated images from 14 methods to demonstrate the erase effectiveness in Figure 16.

Figure 15. The proposed method, DAG, is compared with 13 baselines (11 safe generation methods and 2 bare models) on 10 samples from COCO-1K. Notably, DAG is neither trained nor optimized using MS-COCO dataset, yet it maintains strong concept-specificity.

Figure 16. The proposed method, DAG, is compared with 13 baselines (11 safe generation methods and 2 bare models) on 4 samples from I2P-sexual dataset.