Efficient Fine-tuning and Content Suppression for Pruned Diffusion Models

Reza Shirkavand[1], Peiran Yu[2], Shangqian Gao[3], Gowthami Somepalli[1], Tom Goldstein[1], Heng Huang[1]

1- Department of Computer Science, University of Maryland - College Park

2- Department of Computer Science, University of Texas Arlington

1- Department of Computer Science, Florida State University

DEPARTMENT OF
COMPUTER SCIENCE

Effect of Pruning and Distillation

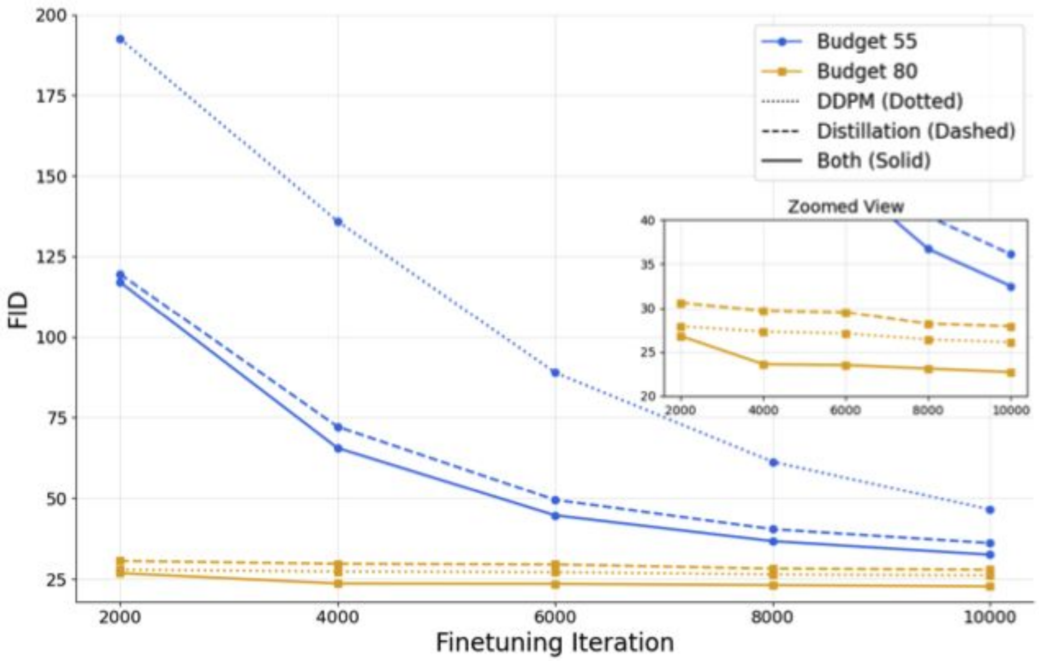


Fig. 1: Effect of Distillation

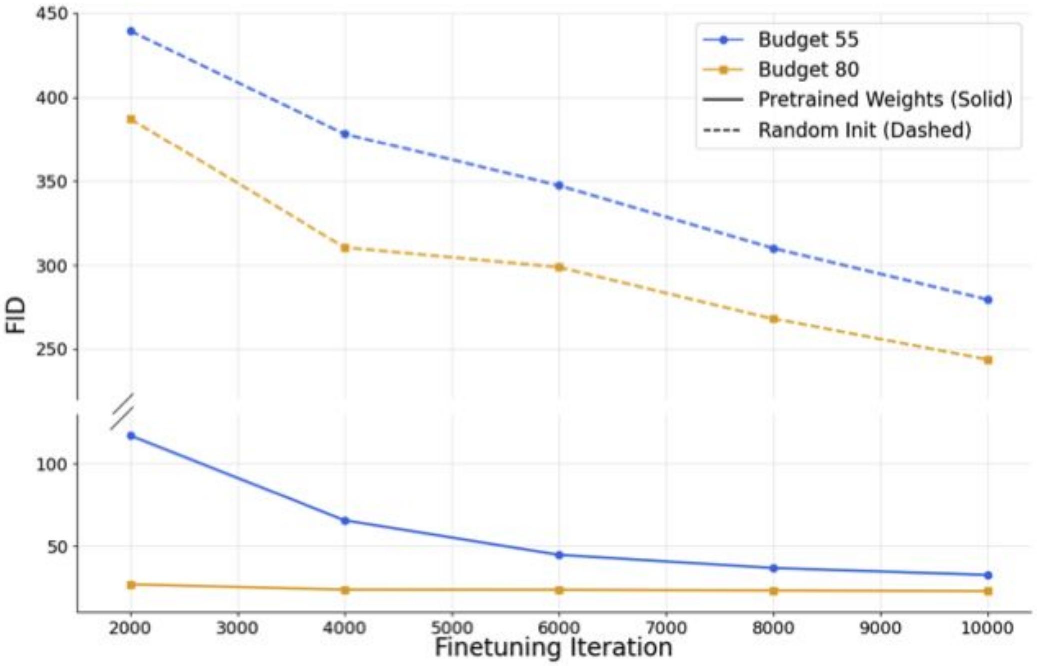


Fig. 2: Effect of Pruning

Effect of Distillation

- Distillation Works but it could be a double-edged sword.

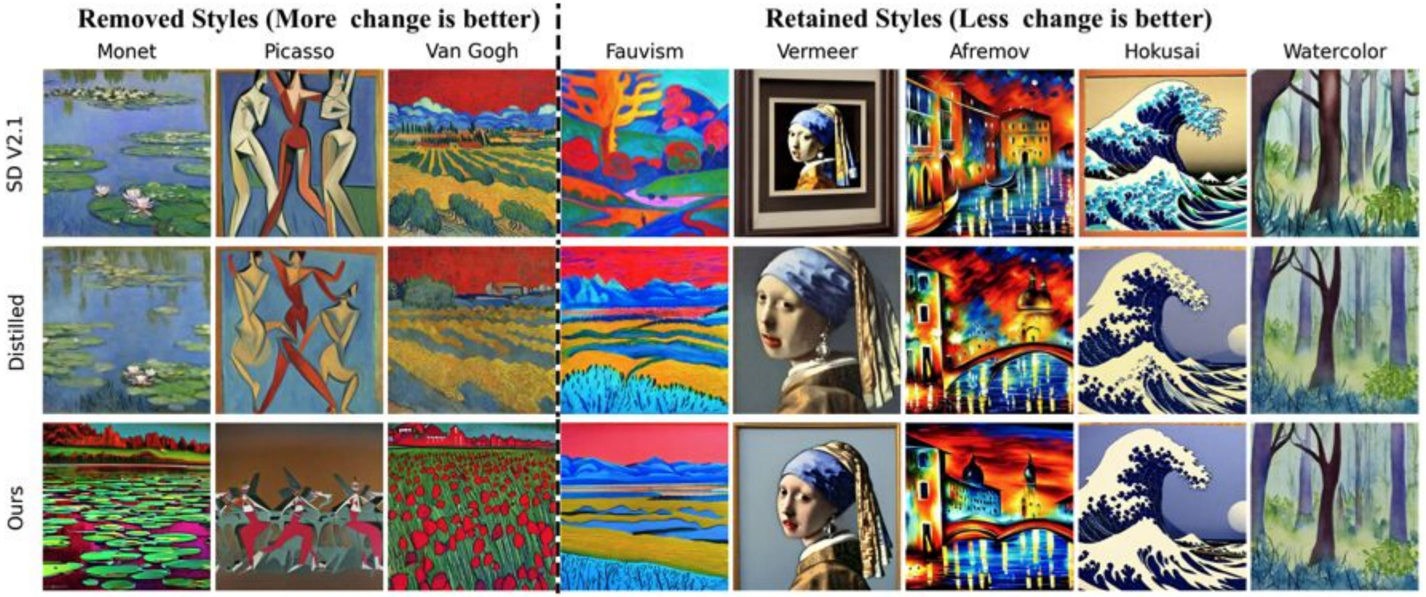


Figure 2. Comparison of generative quality and style adherence: **Row 1:** The original Stable Diffusion 2.1 model. **Row 2:** A pruned version fine-tuned with 20,000 iterations of combined DDPM and distillation loss. **Row 3:** A pruned version fine-tuned with 20,000 iterations of our proposed bilevel fine-tuning approach, removing styles of Van Gogh, Monet, and Picasso. Our bilevel method is successful in retaining generative quality and style diversity while suppressing undesirable concepts. See the Appendix C.2 for prompts used.

Preliminary

- Pruning objective:

$$\min_{\theta_{\text{pruned}}} |L(\theta_{\text{pruned}}) - L(\theta)|, \quad \text{s.t.} \quad \|\theta_{\text{pruned}}\|_0 \leq R,$$

- Distillation objective:

$$\mathcal{L}_D^{\text{Out-KD}} = \mathbb{E}_{x_0,\epsilon,t} \|\epsilon_T(x_t, t, c) - \epsilon_S(x_t, t, c)\|^2,$$
$$\mathcal{L}_D^{\text{Feat-KD}} = \sum_i \mathbb{E}_{x_0,\epsilon,t} \|\epsilon_T^i(x_t, t, c) - \epsilon_S^i(x_t, t, c)\|^2,$$

- Concept unlearning:

$$\min_{\theta_{CU}} \mathbb{E}_{x_0,\epsilon,t,c,c'} \|\epsilon_\theta(x_t, t, c') - \epsilon_{\theta_{CU}}(x_t, t, c)\|^2,$$

$$D_{KL}(p_\theta(x|\bar{c}) \parallel p_{\theta_{CU}}(x|\bar{c})) \approx 0,$$

- Fine-tuning objective of the pruned model:

$$\min_{\theta_{\text{pruned}}} \mathcal{L}^{\text{ft}} := \mathcal{L}^{\text{Diff}} + \lambda^{\text{OutKD}} \mathcal{L}^{\text{OutKD}} + \lambda^{\text{FeatKD}} \mathcal{L}^{\text{FeatKD}}$$

Two-Stage Approach

- **Simple Approach:** Two stage pipeline.
- Assuming the fine-tuning process yields:

$$\hat{\theta} \in \text{argmin}_{\theta_{\text{pruned}}} \mathcal{L}^{\text{ft}}$$

- We initialize with

$$\theta_{CU} \quad \theta' \in \text{argmin}\theta_{\text{CU}} \mathcal{L}^{\text{CU}}$$

- This can be suboptimal.

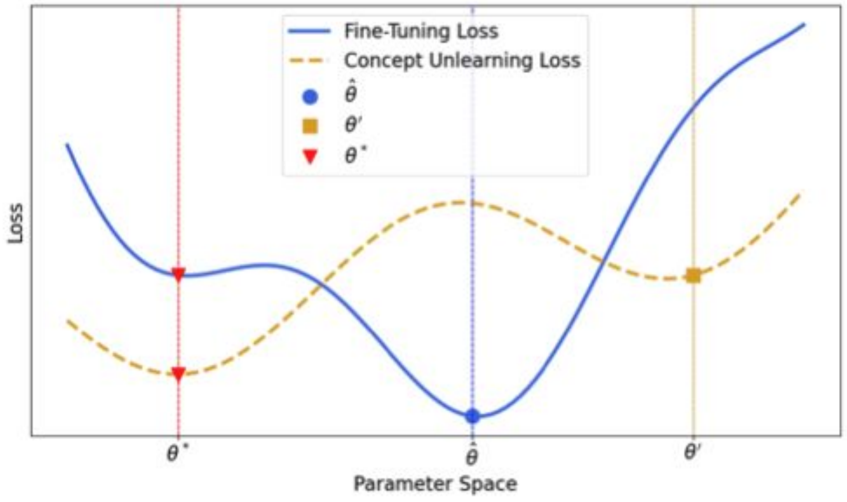


Figure 3. Why can a two-stage approach (fine-tuning followed by forgetting) be suboptimal? If fine-tuning yields $\hat{\theta}$, initializing the concept unlearning parameters with $\hat{\theta}$ and optimizing the concept unlearning loss (Eq. (5)) results in θ', which is suboptimal for both fine-tuning the pruned model and for concept unlearning. In contrast, our bilevel method, defined in Eq. (9), produces the optimal solution θ^*, achieving better performance for both fine-tuning and unlearning.

Our Bilevel Approach

- **Our Approach:** Bilevel optimization.

$$\min_{\theta_{\text{pruned}}} \mathbb{E}_{x_0,\epsilon,t,c,c'} \|\epsilon_\theta(x_t, t, c') - \epsilon_{\theta_{\text{pruned}}}(x_t, t, c)\|^2,$$

$$s.t.\ \theta_{\text{pruned}} \in \text{argmin} \mathcal{L}^{\text{ft}}.$$

- Equivalent to this objective:

$$\min_{\theta_{\text{pruned}}} \mathbb{E}_{x_0,\epsilon,t,c,c'} \|\epsilon_\theta(x_t, t, c') - \epsilon_{\theta_{\text{pruned}}}(x_t, t, c)\|^2,$$

$$s.t.\ \mathcal{L}^{\text{ft}}(\theta_{\text{pruned}}) - \inf_{\vartheta} \mathcal{L}^{\text{ft}}(\vartheta) \leq 0.$$

- Applying a penalty method:

$$\min_{\theta_{\text{pruned}}} \mathcal{L}^{\text{penalized}}(\theta_{\text{pruned}}),$$

$$\begin{aligned}&\mathcal{L}^{\text{penalized}}(\theta_{\text{pruned}}) := \\ &\mathbb{E}_{x_0,\epsilon,t,c,c'} \|\epsilon_\theta(x_t, t, c') - \epsilon_{\theta_{\text{pruned}}}(x_t, t, c)\|^2 \\ &+ \lambda \left(\mathcal{L}^{\text{ft}}(\theta_{\text{pruned}}) - \inf_{\vartheta} \mathcal{L}^{\text{ft}}(\vartheta) \right)\end{aligned}$$

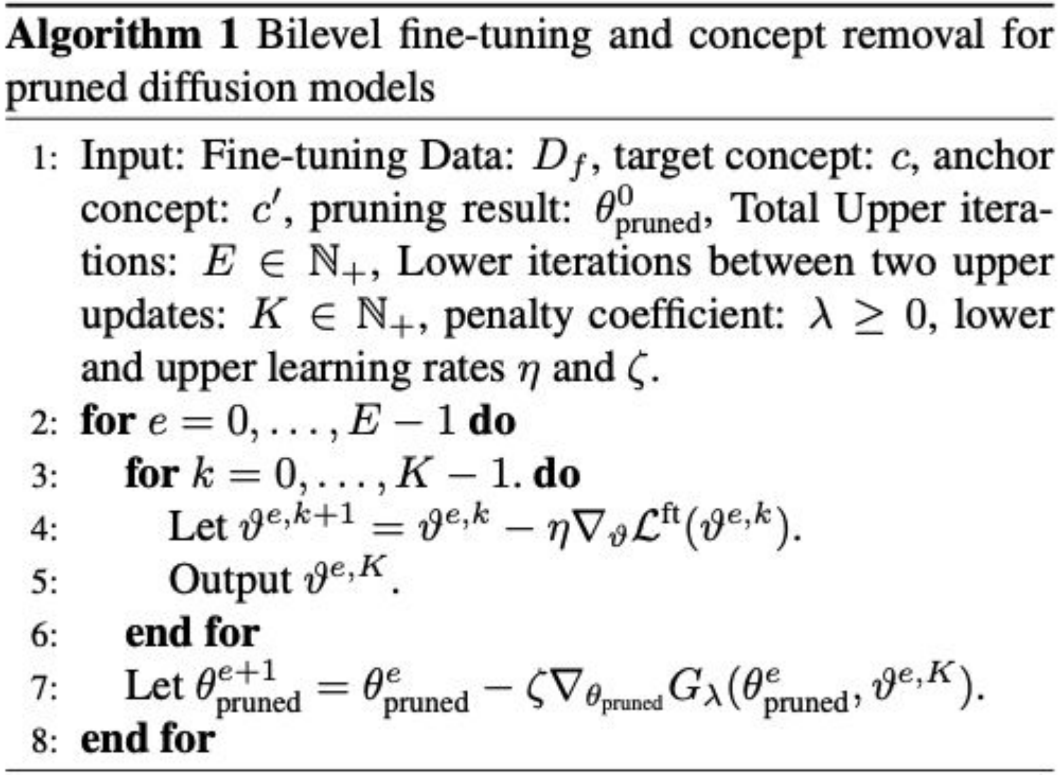
$$\Longrightarrow$$

$$\min_{\theta_{\text{pruned}}} \max_{\vartheta} G_\lambda(\theta_{\text{pruned}}, \vartheta),$$

$$\begin{aligned}&G_\lambda(\theta_{\text{pruned}}, \vartheta) := \\ &\mathbb{E}_{x_0,\epsilon,t,c,c'} \|\epsilon_\theta(x_t, t, c') - \epsilon_{\theta_{\text{pruned}}}(x_t, t, c)\|^2 \\ &+ \lambda \left(\mathcal{L}^{\text{ft}}(\theta_{\text{pruned}}) - \mathcal{L}^{\text{ft}}(\vartheta) \right).\end{aligned}$$

Our Bilevel Approach

- **Our algorithm**



Algorithm 1 Bilevel fine-tuning and concept removal for pruned diffusion models

```
Input: Fine-tuning Data: $D_f$, target concept: $c$, anchor concept: $c'$, pruning result: $\theta^0_{\text{pruned}}$, Total Upper iterations: $E \in \mathbb{N}_+$, Lower iterations between two upper updates: $K \in \mathbb{N}_+$, penalty coefficient: $\lambda \geq 0$, lower and upper learning rates $\eta$ and $\zeta$.
for $e = 0, \dots, E - 1$ do
    for $k = 0, \dots, K - 1.$ do
        Let $\vartheta^{e,k+1} = \vartheta^{e,k} - \eta \nabla_\vartheta \mathcal{L}^{\text{ft}}(\vartheta^{e,k})$.
        Output $\vartheta^{e,K}$.
    end for
    Let $\theta^{e+1}_{\text{pruned}} = \theta^e_{\text{pruned}} - \zeta \nabla_{\theta_{\text{pruned}}} G_\lambda(\theta^e_{\text{pruned}}, \vartheta^{e,K})$.
end for
```

Results

- **Style Removal**

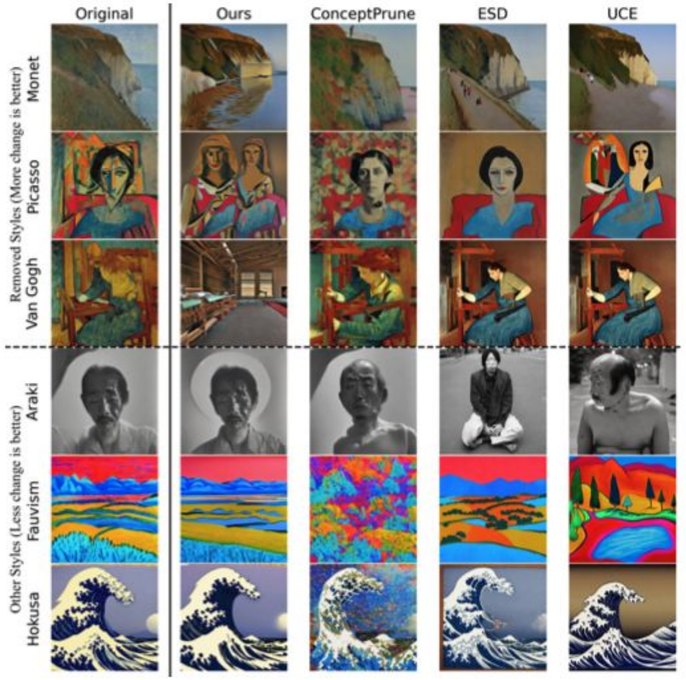
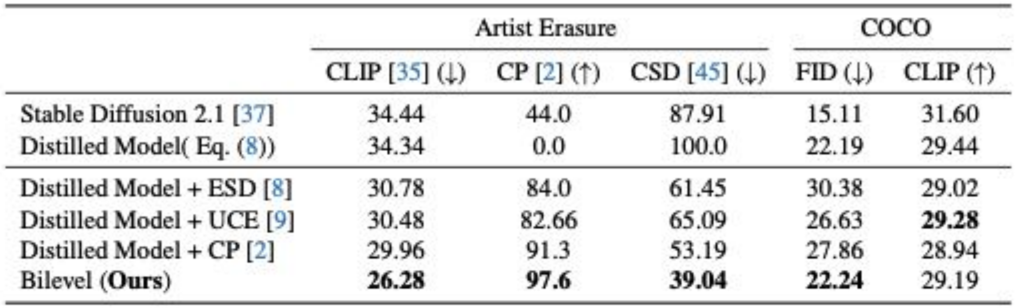

Original
Ours
ConceptPrune
ESD
UCE
Removed Styles (More change is better)
Monet
Picasso
Van Gogh
Other Styles (Less change is better)
Araki
Fauvism
Hokusa

Figure 6. Quantitative results demonstrate the effectiveness of removing the styles of three artists—Monet, Picasso, and Van Gogh—from the pruned model. Our method not only successfully eliminates the target styles completely but also generates other non-removed styles more effectively than the baselines. Original refers to the pruned model that is fine-tuned using only Eq. (8).



	Artist Erasure			COCO	
	CLIP [35] (↓)	CP [2] (↑)	CSD [45] (↓)	FID (↓)	CLIP (↑)
Stable Diffusion 2.1 [37]	34.44	44.0	87.91	15.11	31.60
Distilled Model(Eq. (8))	34.34	0.0	100.0	22.19	29.44
Distilled Model + ESD [8]	30.78	84.0	61.45	30.38	29.02
Distilled Model + UCE [9]	30.48	82.66	65.09	26.63	**29.28**
Distilled Model + CP [2]	29.96	91.3	53.19	27.86	28.94
Bilevel (**Ours**)	**26.28**	**97.6**	**39.04**	**22.24**	29.19

Table 1. **Style Removal**: Quantitative results for removing the styles of three artists—Monet, Picasso, and Van Gogh—from the pruned model across 50 prompts for each artist. CP Score [2] penalizes an unlearning method if the model produces images that have a higher clip score to the style prompt than the original model. CSD [45] is a metric specifically designed to measure style similarity. The COCO values demonstrate the model's ability to retain styles and concepts that were not targeted for removal. Our bilevel method effectively removes the target concepts while restoring the generation capabilities of the pruned model.

Results Cont.

- **NSFW Content Removal**

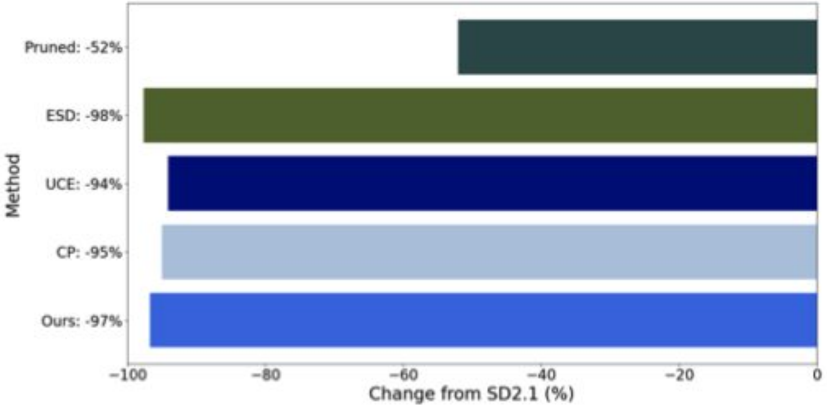
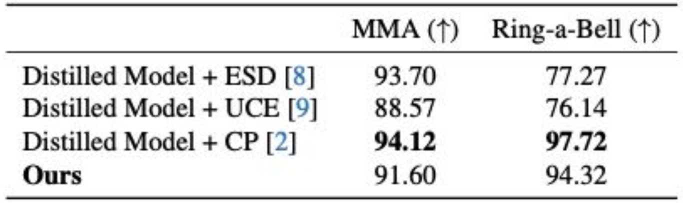
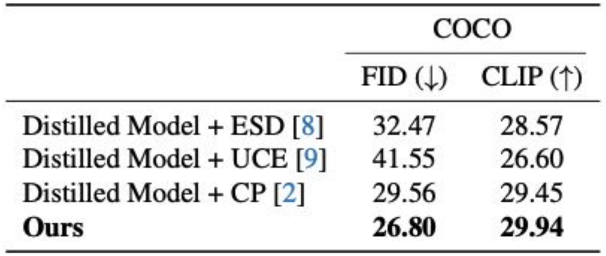


Figure 7. Explicit Content Removal: The values represent the percentage decrease in nudity content in I2P prompts compared to the original SD2.1 model. Pruned baseline performs well as seen by prior work. Our method achieves performance on par with baseline models for NSFW content reduction.



	MMA (↑)	Ring-a-Bell (↑)
Distilled Model + ESD [8]	93.70	77.27
Distilled Model + UCE [9]	88.57	76.14
Distilled Model + CP [2]	**94.12**	**97.72**
Ours	91.60	94.32

Table 2. Comparison of our bilevel method with baseline removal methods on adversarial NSFW prompts: The values indicate the resilience of each method to adversarial prompts. While our method does not outperform all baselines, it demonstrates solid performance on these challenging prompts.



	COCO	
	FID (↓)	CLIP (↑)
Distilled Model + ESD [8]	32.47	28.57
Distilled Model + UCE [9]	41.55	26.60
Distilled Model + CP [2]	29.56	29.45
Ours	**26.80**	**29.94**

Table 3. Quantitative results demonstrating the model's ability to retain styles and concepts that were not targeted for removal by the NSFW removal method. Our bilevel method does not impact the generation capabilities of the model.