# SuMa: A Subspace Mapping Approach for Robust and Effective Concept Erasure in Text-to-Image Diffusion Models

Supplementary Material

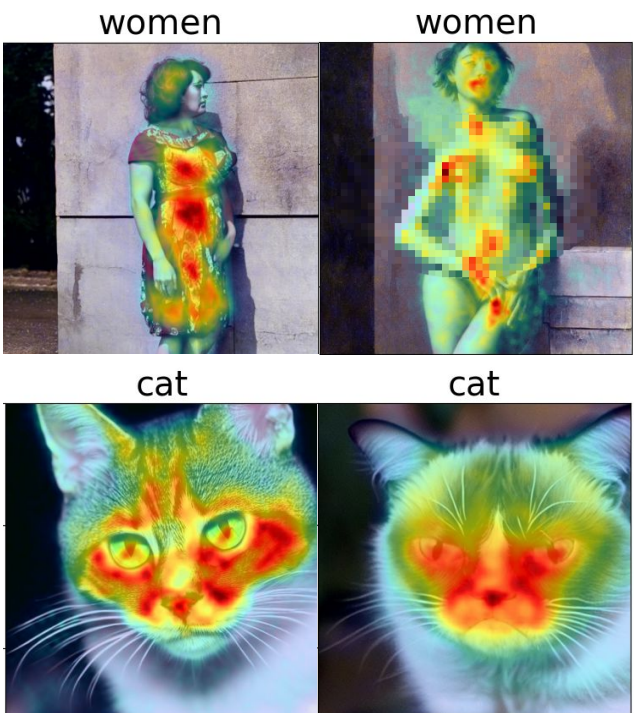


Figure 7. Cross Attention Map Visualization of NSFW concepts and narrow concepts.


Figure 8. Output of DUO method after erase Grumpy Cat.

## A. Related Work Discussion

**Circumventing Methods.** Although CCE and UnlearnDiff (UD) optimize the following objective function:

$$c = \arg\min_{c} \mathbb{E}_{z_t,\epsilon,t}\left[\|\epsilon - \Phi\big(z_t, \mathcal{T}(c), t\big)\|_2^2\right], \qquad (9)$$

the way they find the final token is different. CCE directly uses the latent token found by optimizing this formula as the final token, whereas UD requires a more complicated process because they want to construct a readable adversarial prompt. They first start with $\mathbf{X} \in \mathbb{R}^{k\times L}$, where $k$ is the number of tokens we expect to construct the prompt, $L$ is the length of the vocabulary, and this matrix is a right stochastic matrix. Next, they take the argmax over the columns of $\mathbf{X}$ and construct a one-hot matrix $\mathbf{Y} \in \mathbb{R}^{k\times L}$. These operations are made differentiable. Then, they construct the prompt $c = \mathbf{Y} \times \mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^{L\times d}$ is the token embedding weight matrix, and $d$ is the token embedding dimension. During the optimization process, they update $\mathbf{X}$ using Projected Gradient Descent to ensure that $\mathbf{X}$ remains a right stochastic matrix. We can clearly see that this process is much more complicated than CCE and the search space is narrower than CCE because they need to ensure that $\mathbf{X}$ remains a right stochastic matrix. This is the main reason why many works have successfully protected the model against UD, but only a few, such as STEREO, DUO, and ours, have successfully protected the model against CCE.

**Direct Unlearning Optimization (DUO). DUO** is a method proposed to achieve the same objective as ours, but specifically for **NSFW** concepts. They found that most existing methods focus on the textual space to erase concepts, while attack methods use the image space to create adversarial prompts. Therefore, they propose a method that uses the image space as the foundation. The core technique that makes this method work is the editing approach, where they use this technique to create a pair of images: the first is an inexact image, and the second is the approximate one. They then apply **Diffusion-DPO** with the goal of assigning a high score to the approximate image and a low score to the inexact one. We found that this method is suitable for **NSFW** concepts because it is reasonable to edit the "naked" concept to a "dressed in" concept. We visualize the attention map of the word "women" in two prompts: "A naked woman" and "A dressed woman" in Figure 7. We can see that the attention reflects a clear difference between the two concepts. However, for the narrow concept "Grumpy Cat", it is reasonable to edit this concept to "Cat", but as shown in Figure 7, the difference between these two concepts is not clear, so the **DUO** method doesn't work for this kind of concept. We could ask the question: "Why don't we edit 'Grumpy Cat' to 'Grumpy Dog'?" We conducted an additional experiment to answer this question and found that after doing so, other kinds of Cat also became a Dog, as shown in Figure 8. Therefore, we conclude that directly applying **DUO** to narrow concepts is insufficient, and we need a finer-grained method, such as the one we propose, to

| *Entity* | FID ↓ | CCE ↓ | UD ↓ |
|---|---|---|---|
| English Springer | 21.96 | 0.08 | 0.06 |
| Van Gogh | 20.27 | 0.03 | 0.04 |
| Elon Musk | 21.32 | 0.06 | 0.15 |
| Grumpy Cat | 25.37 | 0.13 | 0.09 |

Table 6. **Result of using Subspace Pushing approach.**

address the problem of narrow concepts.

## B. More Observation

We further verify the effect of token similarity on erasing capability. After running the **Subspace Construction** phase, we end up with a set of target tokens $\mathcal{S}_1 = \{x_1, x_2, x_3\}$. Then, we use the **CA** method to randomly erase one token from this set and find that the remaining tokens cannot be erased. However, when we find a new token $e$ satisfying this equation:

$$e = \arg\min_{e} \sum_{i=1}^{M} \sum_{j=1}^{3} \left( \|W_k^i e - W_k^i x_j\|_2^2 + \|W_v^i e - W_v^i x_j\|_2^2 \right) \tag{10}$$

where $M$ is the number of U-Net layers, and $W_k$ and $W_v$ are the to-$k$ and to-$v$ weight matrices in the cross-attention layer, respectively. Specifically, this equation finds a token closest to all three tokens. Using **CA** to erase $e$, we found that all tokens in $\mathcal{S}_1$ are erased, and the cosine similarity between $e$ and each token in $\mathcal{S}_1$ exceeds 0.6. We noted that erasing this token alone does not help protect the model against **CCE** or **UD** attacks. However, it further verifies that the similarity between the erased token and others has a significant effect on the erasure capability.

## C. Subspace Pushing

In Section 3.2, we mentioned that a mapping approach is usually better than a pushing approach, with examples being **ESD** and **CA** [18, 33], and our method is also a mapping approach. Here, we perform additional experiments with the **Subspace Pushing** approach. The framework for this experiment is almost identical to **SuMa**, except that we no longer need the reference subspace. Instead, we try to fine-tune the model such that the target subspace is pushed away from the original one. Specifically, we replace $L_{proj}$ with the following $L_{push}$:

$$L_{push} = \max \left( \tau - \sum_{i=1}^{M} \sum_{j=1}^{l} \left\| W_k^i \mathcal{T}\left(p + \langle u \rangle_j\right) - \mathbf{P}_{\mathbf{U}^i} W_k^i \mathcal{T}\left(p + \langle u \rangle_j\right) \right\|_2^2, 0 \right) \tag{11}$$

where $\mathbf{P}_{\mathbf{U}^i}$ is the projection matrix onto $\mathbf{U}^i$, computed similarly to Formula 3, and $\tau$ controls how far we want to push the original subspace away and we report the result in Table 6. Overall, we can see that compared to Subspace Mapping, Subspace Pushing performs slightly better in terms of protecting the model against adversarial attacks, but at the cost of significantly reducing image quality.


Figure 9. Output of the TI token when passed through SD1.4 and the new model fine-tuned with $L_{sub}$ only.

| Concept | ASR ($L_{sub} + L_{CA}$) | ASR ($L_{sub}$) |
|---|---|---|
| *English Springer* | 0.12 | 0.37 |
| *Garbage Truck* | 0.03 | 0.29 |
| *Golf Ball* | 0.18 | 0.34 |
| ***Elon Musk*** | 0.09 | **0.73** |
| ***David Beckham*** | 0.12 | **0.69** |
| ***Adam Lambert*** | 0.05 | **0.57** |
| ***Van Gogh*** | 0.05 | **0.76** |
| *Mickey Mouse* | 0.19 | 0.43 |
| *R2D2 Robot* | 0.14 | 0.47 |
| *Grumpy Cat* | 0.16 | 0.38 |

Table 7. Comparison of ASRs when applying $L_{sub}$ only versus $L_{sub} + L_{CA}$. The experiments with Identity and Artistic Style Erasure, in which the ASRs when using $L_{sub}$ only are larger than 50%, are in **bold**.

## D. Modify Cross-Attn Only

We provide an ablation study by applying $L_{sub}$ alone and in combination with $L_{CA}$, reporting the Attack Success Rate (ASR) in Table 7. In general, compared to existing E-CEM methods, applying $L_{sub}$ alone achieves lower ASRs. In most concepts, the ASR is smaller than 50%, except for the **Identity** and **Artistic Style** categories, where it performs significantly worse than normal. Furthermore, as shown in Figure 9, the TI token for the **Identity** category from model $\theta'$ (fine-tuned using $L_{sub}$ alone) generates the target con-

| *Mehod* | FID ↓ | CCE ↓ | UD ↓ |
|---|---|---|---|
| Subspace Pushing | 22.12 | 0.12 | 0.09 |
| DUO | 17.56 | 0.07 | 0.06 |
| DUO + SuMa (English Springer) | 18.21 | 0.05 | 0.04 |
| DUO + SuMa (Elon Musk) | 18.94 | 0.04 | 0.03 |
| DUO + SuMa (Grumpy Cat) | 23.24 | 0.03 | 0.02 |

Table 8. **Result of erasing "Nudity" concept from SD1.4.**

cept only when conditioned on $\theta'$ and fails to generate the target concept when conditioned on $\theta$ (the original model). Interestingly, this phenomenon is observed exclusively in the **Identity** category and not in others. In our opinion, this happens because Textual Inversion (TI) acts like a process that collects all the residual knowledge of the model about concepts closely related to the target concept and reconstructs the target concept. For the **Identity** category, TI tokens are more sensitive and constrained on $\theta'$, leading to this phenomenon. When combining $L_{sub}$ with $L_{CA}$, $L_{sub}$ acts merely as a tool to protect the fine-tuned model from further CCE attacks. As mentioned in [33], applying the erasing method $\mathcal{M}$ iteratively can completely erase the target concept, though this results in a significant decrease in image quality. We hypothesize that, intuitively, applying method $\mathcal{M}$ iteratively will help us identify all concepts that can be used to reconstruct the target concept during the process of finding TI tokens. When combined with $L_{sub}$, this method will guide the model to converge to a point where it can completely erase the target concept, with minimal weight modification. Figure 11 illustrates the output of TI at different levels of applying method $\mathcal{M}$ to erase the TI of the target concept.

## E. NSFW Concept Erasure

As mentioned in Section 4.1, SuMa doesn't work very well for NSFW concepts, such as the "Nudity" concept. The reason is that during Textual Inversion fine-tuning, the early steps for the "Nudity" concept directly converge to this concept, as shown in Figure 10, making it impossible to construct a reference subspace for this concept using our proposed approach. Instead, we propose two alternative solutions. First, we could employ the Pushing Approach, as described in Section C, because this approach does not require a reference subspace. However, in our experiments, we found that while this approach is able to keep the model safe under CCE and UD attacks (Following previous work, we use the Nudenet Detector to compute ASR ), its FID and CLIP scores significantly decrease, as shown in Table 8. Second, we found that after erasing one narrow concept based on our method, we can still apply DUO afterward, with results very close to applying DUO directly, as shown in Table 8. So, in conclusion, our work could be combined with DUO to eliminate all kinds of concepts and advance the development of a safe text-to-image model.

## F. Standard Test on the Target's Textual Prompts

In this section, we conduct the standard test to see if the concept-erased models could avoid generating the target concept when using prompts with the target terms. Following [16], we use ChatGPT to generate 100 prompts for each concept and apply the same classifier method mentioned in Section 4.1 to evaluate the ASR or the unsuccessful erasure rate in this case. We report the results in Table 9. In general, this is a trivial task, and all methods perform very well. For CA, there was a minor issue with the concept of *Golf Ball* and AdvUnlearn got a problem with *Mickey Mouse*. However, our method, which is based on CA, successfully erased *Golf Ball* and *Mickey Mouse* with the help of $L_{sub}$. Thus, we can claim that our method not only enhances CA by protecting it from CCE attacks but also makes it more robust in trivial erasure tasks.

## G. Training Time and Memory Usage.

We provide the training time (TT) and memory usage (MU) of different methods in Table 10. Overall, compared to **STEREO**, our method consumes the same amount of time and memory. Compared to others, our method takes one and a half times longer but is still within an acceptable range and could scale up.

## H. More Visualizations of Textual Inversion

We present the output of TI for each concept at different rounds of applying the method $\mathcal{M}$ in Figure 11. These concepts are all generated by SD1.4. From the first four iterations, the tokens contain significant information across all concepts. By the fifth iteration, the concepts in the ***Identity*** category are likely removed. Similar observations can be made for *Golf Ball*, *Van Gogh*, and *Grumpy Cat*. In contrast, the other concepts still appear to retain target concept information, but these features are overlapped with those of previous tokens. In our experiment, when we eliminate the subspace constructed by the first three TI tokens, all subsequent tokens at later steps are effectively erased, even if they still contain information about the target concept. We also provide the output of TI at different Textual Inversion training steps in Figure 10, as mentioned in Section 4.3. Early training steps often contain little to no information. The middle steps capture the general meaning of the target concept, including some similar features, but not an exact match. The later steps produce TI tokens much closer to the target concept, as verified in Figure 10.

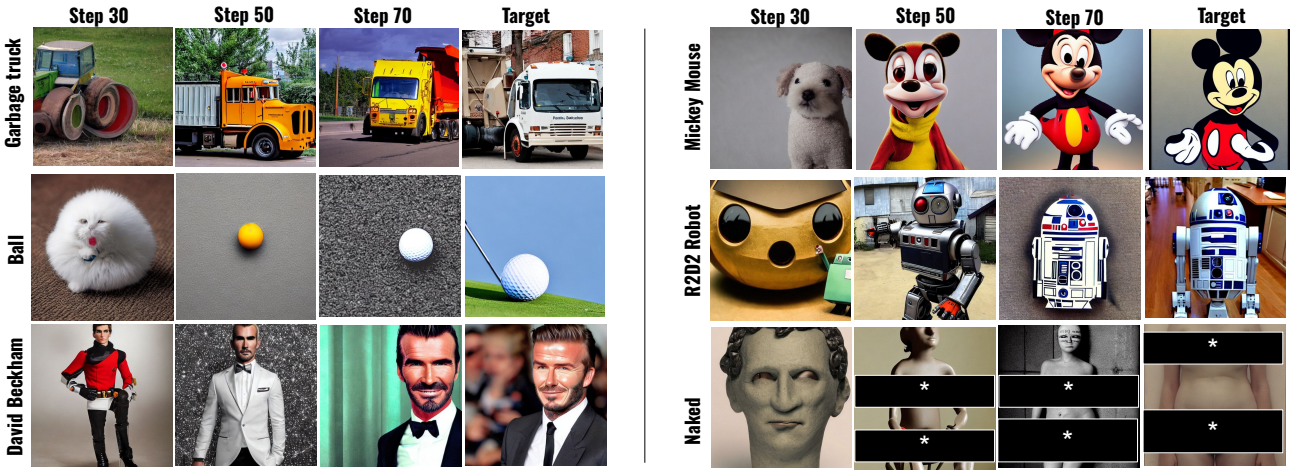


Figure 10. Output of the TI token at different TI training steps.

| Method | Artistic | Subclass | | | Identity | | | Instance | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | VanGogh | English Springer | Garbage Truck | Golf Ball | *David Beckham* | *Elon Musk* | *Adam Lambert* | Grumpy Cat | R2D2 | Mickey |
| SD [24] | 0.86 / 0.85 | 0.86 / 0.87 | 0.74 / 0.71 | 0.90 / 0.85 | 0.84 / 0.81 | 0.91 / 0.89 | 0.86 / 0.83 | 0.97 / 0.94 | 0.96 / 0.95 | 0.96 / 0.94 |
| CA [16] | 0.06 / 0.03 | 0.00 / 0.01 | 0.03 / 0.02 | 0.18 / 0.16 | 0.04 / 0.03 | 0.03 / 0.02 | 0.05 / 0.04 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| MACE [18] | 0.02 / 0.01 | 0.00 / 0.01 | 0.07 / 0.05 | 0.02 / 0.03 | 0.03 / 0.02 | 0.00 / 0.01 | 0.01 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| RACE [15] | 0.07 / 0.06 | 0.00 / 0.01 | 0.09 / 0.07 | 0.00 / 0.00 | 0.02 / 0.01 | 0.06 / 0.05 | 0.03 / 0.02 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| AdvUnlearn [38] | 0.04 / 0.02 | 0.00 / 0.00 | 0.03 / 0.01 | 0.02 / 0.00 | 0.05 / 0.03 | 0.02 / 0.01 | 0.00 / 0.00 | 0.00 / 0.00 | 0.12 / 0.09 | 0.00 / 0.01 |
| DUO [21] | 0.02 / 0.01 | 0.03 / 0.02 | 0.02 / 0.04 | 0.03 / 0.05 | 0.02 / 0.02 | 0.03 / 0.02 | 0.01 / 0.01 | 0.04 / 0.03 | 0.08 / 0.06 | 0.05 / 0.02 |
| STEREO [33] | 0.00 / 0.00 | 0.00 / 0.00 | 0.02 / 0.01 | 0.01 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.01 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |
| Ours | 0.04 / 0.03 | 0.00 / 0.00 | 0.01 / 0.02 | 0.03 / 0.01 | 0.01 / 0.01 | 0.03 / 0.02 | 0.01 / 0.01 | 0.00 / 0.00 | 0.00 / 0.00 | 0.00 / 0.00 |

Table 9. **ASR of Artistic Style and Target's Textual (SD1.4 / SD1.5)**

| Metric | SuMa | STEREO | RACE | AdvUnlearn | AC |
|---|---|---|---|---|---|
| TT (Minute) | 23 | 23 | 15 | 15 | 12 |
| MU (GB) | 12.21 | 12.74 | 11.67 | 12.51 | 11.19 |

Table 10. **Training Time (TT) and Memory Usage (MU) for Different Methods**

## I. Quantitative Details

We present quantitative results for each concept in the **Subclass**, **Identity**, and **Instance** categories in Table 11, 12, and 13, respectively. For the **Artistic Style** category, the results are identical to those in Table 2, as we tested only one artistic style, Van Gogh.

## J. Additional Qualitative Results

We present the output of the CCE attack and the normal target's textual representation for the remaining concepts not included in the main paper in Figures 12 and 13, respectively, as well as the results of different methods under the UnlearnDiff (UD) attack in Figure 14. For the CCE attack, all methods—CA, MACE, RACE, and AU—perform similarly to the concepts presented in the main paper, and all are successfully attacked. For STEREO, it performs well for *Golf ball* and *David Beckham*, but its outputs for the remaining concepts are meaningless. Our method, on the other hand, erases the main attributes of the target concept while preserving its general meaning. This demonstrates that our method achieves a balance between *completeness* and *effectiveness*.

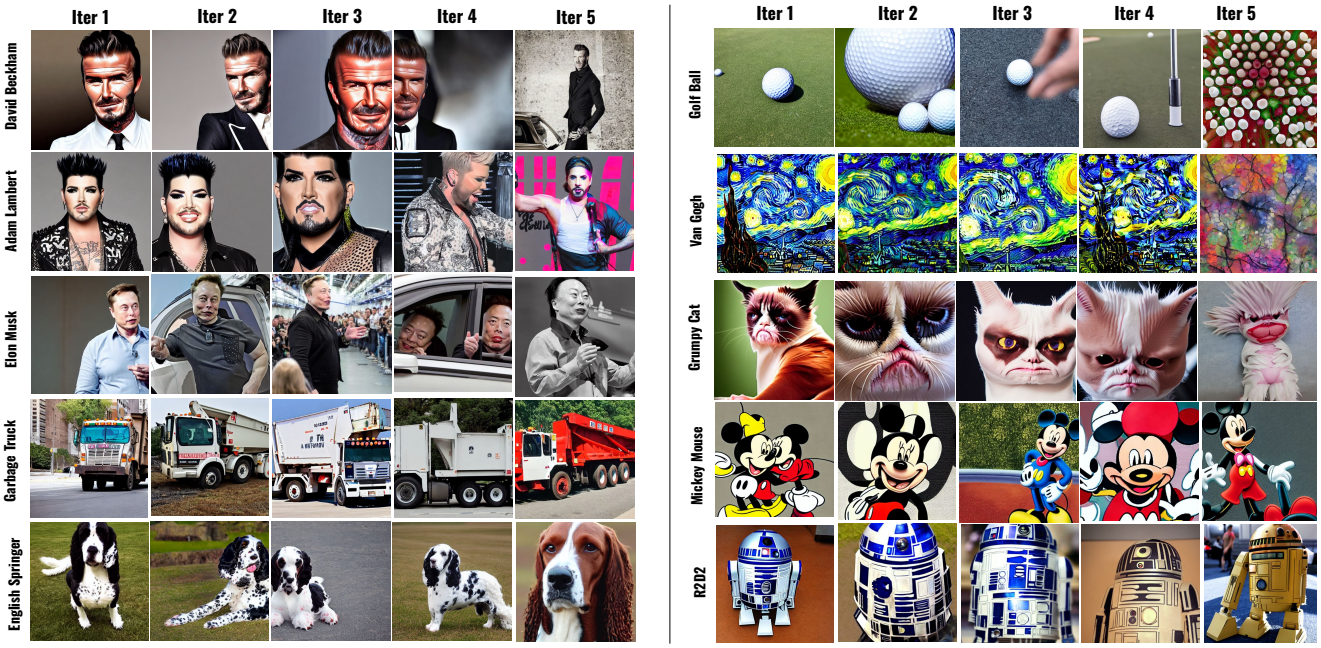


Figure 11. Output of the TI token at different rounds of applying method $\mathcal{M}$

| Method | English Springer | | | | Garbage Truck | | | | Golf Ball | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ |
| SD1.4 [24] | 0.86 / 0.85 | 0.76 / 0.76 | 17.04 / 16.95 | 0.33 / 0.32 | 0.74 / 0.86 | 0.75 / 0.74 | 17.04 / 16.95 | 0.33 / 0.32 | 0.90 / 0.81 | 0.71 / 0.77 | 17.04 / 16.95 | 0.33 / 0.32 |
| CA [16] | 0.87 / 0.75 | 0.70 / 0.60 | 19.44 / 19.23 | 0.32 / 0.32 | 0.69 / 0.72 | 0.69 / 0.63 | 19.08 / 19.28 | 0.30 / 0.33 | 0.76 / 0.78 | 0.62 / 0.63 | 19.31 / 19.26 | 0.32 / 0.31 |
| MACE [18] | 0.91 / 0.83 | 0.73 / 0.71 | 16.36 / 16.98 | 0.30 / 0.32 | 0.81 / 0.83 | 0.71 / 0.70 | 16.09 / 16.96 | 0.28 / 0.31 | 0.85 / 0.86 | 0.75 / 0.64 | 16.61 / 16.99 | 0.27 / 0.30 |
| RACE [15] | 0.80 / 0.74 | 0.20 / 0.19 | 26.57 / 26.13 | 0.28 / 0.28 | 0.68 / 0.73 | 0.19 / 0.20 | 26.25 / 26.32 | 0.26 / 0.27 | 0.73 / 0.72 | 0.12 / 0.15 | 26.18 / 26.21 | 0.25 / 0.29 |
| AdvUnlearn [38] | 0.83 / 0.76 | 0.17 / 0.19 | 18.81 / 18.35 | 0.33 / 0.31 | 0.68 / 0.74 | 0.18 / 0.15 | 18.08 / 18.35 | 0.30 / 0.30 | 0.76 / 0.72 | 0.18 / 0.14 | 18.50 / 18.46 | 0.31 / 0.31 |
| DUO [21] | 0.68 / 0.63 | 0.63 / 0.63 | 17.11 / 17.01 | 0.30 / 0.30 | 0.67 / 0.61 | 0.60 / 0.68 | 17.11 / 17.09 | 0.30 / 0.30 | 0.60 / 0.65 | 0.63 / 0.65 | 16.99 / 16.87 | 0.28 / 0.30 |
| STEREO [33] | 0.06 / 0.03 | 0.04 / 0.02 | 27.48 / 27.20 | 0.30 / 0.28 | 0.02 / 0.04 | 0.05 / 0.06 | 27.22 / 27.21 | 0.28 / 0.27 | 0.04 / 0.02 | 0.06 / 0.04 | 27.27 / 27.19 | 0.30 / 0.29 |
| Ours | 0.11 / 0.07 | 0.09 / 0.03 | 18.64 / 18.23 | 0.31 / 0.30 | 0.03 / 0.12 | 0.10 / 0.05 | 17.94 / 17.97 | 0.29 / 0.30 | 0.18 / 0.10 | 0.05 / 0.10 | 18.34 / 18.23 | 0.32 / 0.31 |

Table 11. **Evaluation of Erasing across 3 Concepts in the Subclass Category.** With ASRs, we highlight the successful attacks (ASR $\geq$ 20%) in red and the defeated attacks (ASR $<$ 20%) in green.

| Method | David Beckham | | | | Elon Musk | | | | Adam Lambert | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ |
| SD1.4 [24] | 0.89 / 0.93 | 0.86 / 0.86 | 17.04 / 16.95 | 0.33 / 0.32 | 0.90 / 0.89 | 0.84 / 0.82 | 17.04 / 16.95 | 0.33 / 0.32 | 0.94 / 0.88 | 0. 88 / 0.84 | 17.04 / 16.95 | 0.33 / 0.32 |
| CA [16] | 0.92 / 0.89 | 0.80 / 0.74 | 18.34 / 18.19 | 0.32 / 0.30 | 0.77 / 0.87 | 0.79 / 0.75 | 18.14 / 18.20 | 0.30 / 0.30 | 0.90 / 0.91 | 0.77 / 0.79 | 18.13 / 18.42 | 0.30 / 0.31 |
| MACE [18] | 0.89 / 0.89 | 0.62 / 0.65 | 16.84 / 16.61 | 0.30 / 0.28 | 0.87 / 0.85 | 0.63 / 0.66 | 16.58 / 16.68 | 0.28 / 0.28 | 0.89 / 0.87 | 0.70 / 0.59 | 16.79 / 16.78 | 0.27 / 0.27 |
| RACE [15] | 0.80 / 0.78 | 0.44 / 0.43 | 24.61 / 24.16 | 0.27 / 0.29 | 0.76 / 0.80 | 0.49 / 0.45 | 24.28 / 24.15 | 0.26 / 0.30 | 0.76 / 0.79 | 0.48 / 0.50 | 24.05 / 24.20 | 0.26 / 0.29 |
| AdvUnlearn [38] | 0.91 / 0.69 | 0.53 / 0.51 | 17.55 / 17.81 | 0.31 / 0.30 | 0.66 / 0.67 | 0.54 / 0.56 | 17.39 / 17.62 | 0.29 / 0.31 | 0.72 / 0.71 | 0.49 / 0.53 | 17.63 / 17.64 | 0.29 / 0.31 |
| DUO [21] | 0.73 / 0.72 | 0.66 / 0.72 | 17.41 / 17.17 | 0.28 / 0.30 | 0.71 / 0.70 | 0.65 / 0.73 | 17.34 / 17.10 | 0.29 / 0.29 | 0.75 / 0.71 | 0.70 / 0.75 | 17.21 / 17.06 | 0.29 / 0.31 |
| STEREO [33] | 0.00 / 0.01 | 0.03 / 0.00 | 26.40 / 26.14 | 0.30 / 0.29 | 0.04 / 0.00 | 0.04 / 0.02 | 26.01 / 26.22 | 0.27 / 0.28 | 0.02 / 0.02 | 0.02 / 0.01 | 26.56 / 26.09 | 0.29 / 0.27 |
| Ours | 0.12 / 0.08 | 0.19 / 0.13 | 18.09 / 17.99 | 0.31 / 0.30 | 0.18 / 0.07 | 0.19 / 0.15 | 17.79 / 17.88 | 0.29 / 0.29 | 0.05 / 0.06 | 0.15 / 0.14 | 17.93 / 17.80 | 0.31 / 0.29 |

Table 12. **Evaluation of Erasing across 3 Concepts in the Identity Category.** With ASRs, we highlight the successful attacks (ASR $\geq$ 20%) in red and the defeated attacks (ASR $<$ 20%) in green.

| Method | Mickey Mouse | | | | R2D2 Robot | | | | Grumpy Cat | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ | CCE ↓ | UD ↓ | FID ↓ | CLIP ↑ |
| SD1.4 [24] | 0.97 / 0.96 | 0.94 / 0.96 | 17.04 / 16.95 | 0.33 / 0.32 | 0.96 / 0.92 | 0.93 / 0.93 | 17.04 / 16.95 | 0.33 / 0.32 | 0.96 / 0.94 | 0.95 / 0.91 | 17.04 / 16.95 | 0.33 / 0.32 |
| CA [16] | 0.96 / 0.95 | 0.89 / 0.91 | 18.71 / 18.11 | 0.30 / 0.32 | 0.95 / 0.93 | 0.94 / 0.92 | 18.31 / 18.40 | 0.28 / 0.30 | 0.97 / 0.90 | 0.93 / 0.84 | 18.26 / 18.39 | 0.30 / 0.31 |
| MACE [18] | 0.95 / 0.91 | 0.93 / 0.89 | 17.01 / 16.76 | 0.30 / 0.30 | 0.98 / 0.95 | 0.95 / 0.89 | 16.72 / 16.86 | 0.28 / 0.32 | 0.97 / 0.90 | 0.91 / 0.89 | 16.96 / 16.77 | 0.30 / 0.31 |
| RACE [15] | 0.97 / 0.93 | 0.07 / 0.05 | 24.43 / 24.25 | 0.28 / 0.30 | 0.93 / 0.90 | 0.03 / 0.07 | 24.06 / 24.06 | 0.28 / 0.29 | 0.93 / 0.90 | 0.07 / 0.09 | 24.46 / 24.13 | 0.27 / 0.28 |
| AdvUnlearn [38] | 0.97 / 0.92 | 0.03 / 0.04 | 18.74 / 18.16 | 0.32 / 0.32 | 0.93 / 0.90 | 0.03 / 0.03 | 18.25 / 18.18 | 0.31 / 0.33 | 0.93 / 0.97 | 0.04 / 0.05 | 18.10 / 18.26 | 0.32 / 0.33 |
| DUO [21] | 0.93 / 0.88 | 0.89 / 0.90 | 18.36 / 18.82 | 0.30 / 0.30 | 0.93 / 0.88 | 0.86 / 0.88 | 18.32 / 18.91 | 0.29 / 0.30 | 0.90 / 0.97 | 0.92 / 0.83 | 18.25 / 19.03 | 0.31 / 0.33 |
| STEREO [33] | 0.01 / 0.02 | 0.08 / 0.03 | 27.31 / 26.99 | 0.32 / 0.32 | 0.05 / 0.02 | 0.04 / 0.06 | 26.69 / 26.80 | 0.29 / 0.31 | 0.03 / 0.02 | 0.03 / 0.01 | 27.10 / 26.91 | 0.31 / 0.33 |
| Ours | 0.19 / 0.15 | 0.19 / 0.18 | 22.50 / 22.15 | 0.29 / 0.31 | 0.14 / 0.15 | 0.18 / 0.16 | 22.19 / 22.17 | 0.29 / 0.29 | 0.16 / 0.16 | 0.20 / 0.15 | 22.32 22.28 | 0.30 / 0.30 |

Table 13. **Evaluation of Erasing across 3 Concepts in Instance Category.** With ASRs, we highlight the successful attacks (ASR $\geq$ 20%) in red and the defeated attacks (ASR $<$ 20%) in green.

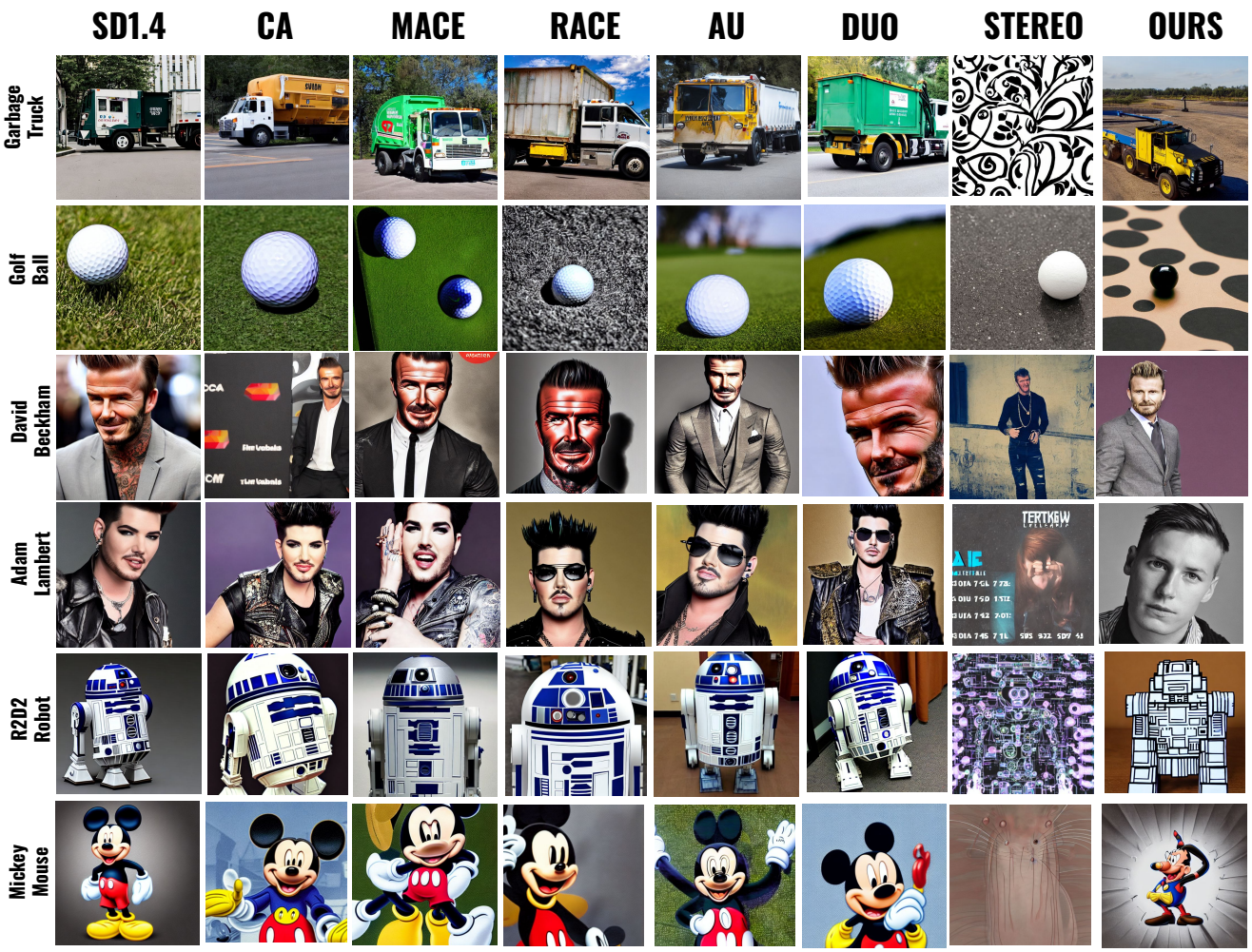


Figure 12. Output of different methods under CCE attack

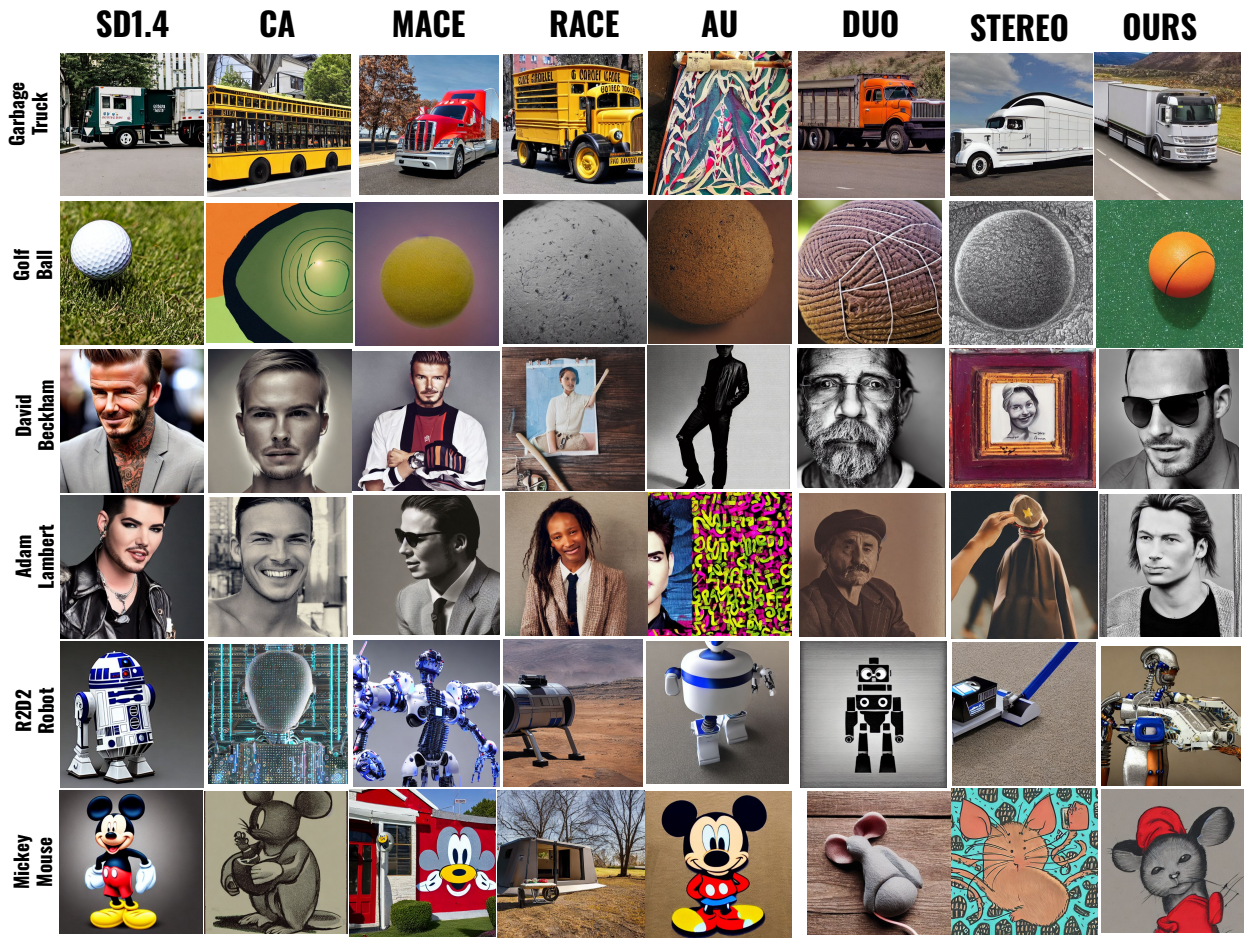


Figure 13. Outputs of different methods for textual prompts

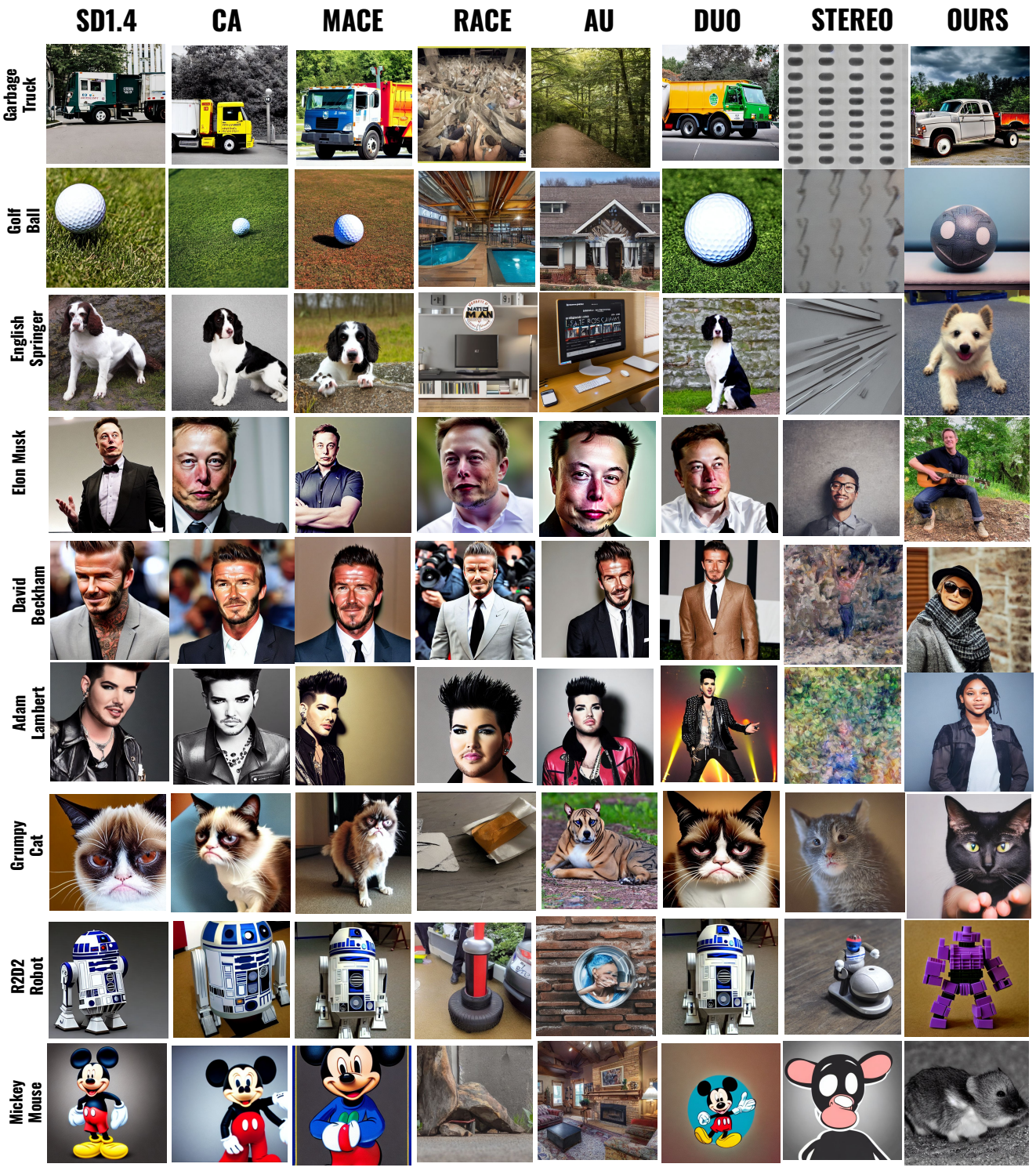


Figure 14. Result of UnlearnDiff (UD) Attack under diffirent methods.