A. Deriving the Closed Form Solution for UCE

Let W_k and W_v be the cross attention weights that project the text embeddings to image space corresponding to keys and values respectively. These are computed fresh at every time step making the computation straightforward, unlike the self-attentions.

Let $\{c_i\}_{i=0}^{m}$ be the embeddings of the text descriptions for the concepts we want to edit. Let $\{c_j\}_{i=0}^{n}$ be the embeddings of the concepts we wish to preserve. Similarly, let $\{v_i^*\}_{i=0}^{m}$ be the target cross-attention outputs that we want the concepts c_i to be steered towards. The sets c_i, v_i^*, and c_j represent the concepts to erase, desired target outputs, and concepts to preserve respectively. We optimize the newly edited weights of the cross-attention value projects W by minimizing the loss function. The same optimization can be adopted to value projection optimization:

$$\mathcal{L} = \sum_{c_i \in E} ||Wc_i - v_i^*||_2^2 + \sum_{c_j \in P} ||Wc_j - W^{old}c_j||_2^2 \tag{A.1}$$

The objective function in Equation A.1 can be solved to arrive at a closed-form solution. We take the derivative of the loss function stated above w.r.t to W and set it to 0.

$$\sum_{c_i \in E} 2(Wc_i - v_i^*)c_i^T + \sum_{c_j \in P} 2(Wc_j - W^{old}c_j)c_j^T = 0$$

$$\sum_{c_i \in E} Wc_ic_i^T - \sum_{c_i \in E} v_i^*c_i^T + \sum_{c_j \in P} Wc_jc_j^T - \sum_{c_j \in P} W^{old}c_jc_j^T = 0$$

$$\sum_{c_i \in E} Wc_ic_i^T + \sum_{c_j \in P} Wc_jc_j^T = \sum_{c_i \in E} v_i^*c_i^T + \sum_{c_j \in P} W^{old}c_jc_j^T$$

$$W\left(\sum_{c_i \in E} c_ic_i^T + \sum_{c_j \in P} c_jc_j^T\right) = \left(\sum_{c_i \in E} v_i^*c_i^T + \sum_{c_j \in P} W^{old}c_jc_j^T\right)$$

To invert the terms on the left-hand side ($\sum_{c_i \in E} c_ic_i^T + \sum_{c_j \in P} c_jc_j^T$), the matrix must have full rank. Adding a preservation term increases the rank by 1. Thus if the number of preservation terms $|P| < d$, where d is the dimension of the text embedding space, the matrix may not have full rank. To ensure the rank condition is satisfied, we introduce d additional preservation terms along the canonical basis directions of the text embedding space. This maintains full rank and enables inversion of the matrix irrespective of the size of P.

$$W = \left(\sum_{c_i \in E} v_i^*c_i^T + \sum_{c_j \in P} W^{old}c_jc_j^T\right)\left(\sum_{c_i \in E} c_ic_i^T + \sum_{c_j \in P} c_jc_j^T\right)^{-1}$$

We optimize both the cross-attention key and value weights using the same principles.

B. UCE Generalizes to TIME

Our method, Unified Concept Editing (UCE), can be viewed as a generalization of the TIME method. As discussed in the methodology section, TIME regularizes the cross-attention weights. With our method, if we do not preserve any specific concepts and only preserve the canonical directions e_j scaled by λ, we get the following closed-form solution:

$$W = \left(\sum_{c_i \in E} v_i^*c_i^T + \lambda \sum_{j=0}^{d} W^{old}e_je_j^T\right)\left(\sum_{c_i \in E} c_ic_i^T + \lambda \sum_{j=0}^{d} e_je_j^T\right)^{-1}$$

Where the canonical directions e_j have outer products $e_ie_j^T$ that are diagonal matrices with only the j^{th} element as 1 and rest 0. Summing all the canonical outer products gives the identity matrix $\mathbb{I}$.

$$W = \left(\sum_{c_i \in E} v_i^*c_i^T + \lambda W^{old}.\mathbb{I}\right)\left(\sum_{c_i \in E} c_ic_i^T + \lambda \mathbb{I}\right)^{-1}$$

This is the closed-form solution for TIME, which regularizes equally across all directions. Our method can be seen as a generalization of TIME that adds preservation across important surrounding concepts, not just the canonical directions. This new formulation with additional explicit preservation is very practical, allowing us to edit multiple concepts with less interference. Our method builds on TIME by allowing the preservation of concepts beyond just the canonical directions. This helps enable editing multiple concepts with less interference.

C. UCE Generalizes to MEMIT

Our method can also be viewed as a generalization of MEMIT. Starting from our objective function in Equation A.1:

$$\mathcal{L} = \sum_{c_i \in E} ||Wc_i - v_i^*||_2^2 + \sum_{c_j \in P} ||Wc_j - W^{old}c_j||_2^2$$

Taking the derivative and setting it to zero gives:

$$\sum_{c_i \in E} 2(Wc_i - v_i^*)c_i^T + \sum_{c_j \in P} 2(Wc_j - W^{old}c_j)c_j^T = 0$$

To ensure full rank, instead of adding canonical directions to complete the rank, we add additional preservations for a plethora of concepts in diffusion vocabulary. We rewrite the equation in a block form by defining $v_j = W^{old}c_j$ and redefining W as $W^{old} + \Delta W$:

$$(W^{old} + \Delta W)(C_iC_i^T + C_jC_j^T) = V_iC_i^T + V_jC_j^T$$

$$W^{old}C_iC_i^T + W^{old}C_jC_j^T + \Delta WC_iC_i^T + \Delta WC_jC_j^T = V_iC_i^T + V_jC_j^T \tag{C.1}$$

Assuming the preservation list contains most concepts the diffusion model knows W^{old} will minimize $||Wc_j - v_j||_2^2$. Taking a derivative and equating to zero, we get $W^{old}C_jC_j^T = V_jC_j^T$. Subtracting this from Equation C.1 gives MEMIT closed form solution:

$$\Delta W(C_iC_i^T + C_jC_j^T) = V_iC_i^T - W^{old}C_iC_i^T$$

With $R = V_i - W^{old}C_i$ and $C_0 = C_jC_j^T$

$$\Delta W = RC_i^T(C_iC_i^T + C_jC_j^T)^{-1}$$

In summary, our method generalizes MEMIT by incorporating additional preservation terms from the diffusion model's vocabulary and solving for the weight update ΔW instead of directly solving for W. This highlights the connection between our approach and existing techniques like MEMIT.

D. Extended Experimental Results

D.1. Erasing Style

We tested the limits of erasing artistic styles using our technique. As shown in Figure D.1, quality for holdout artists declines when erasing over 100 styles, evidenced by the increasing LPIPS after 100 erasures. With 50 or fewer erasures, interference was minimal for non-targeted concepts. Additional qualitative results for our method and baselines are provided in Figures D.4-D.7. The baselines are less effective at removing multiple artists and exhibit greater interference on unerased styles compared to our approach. Our method demonstrates superior erasure while minimizing interference when removing multiple artists. Figure D.8 shows the results of stress testing the limits of artistic style erasure before general art capabilities decline. We observed the model starts to lose artistic nuance in generated outputs after approximately 1000 edits.

An important follow-up question is the minimum number of artists requiring preservation to maintain performance when erasing styles. We analyzed this by testing preservation limits when erasing 10 artists, as shown in Figure D.3. Erasing up to 1500 artists while preserving subsets, we assessed the impact on 100 non-preserved, non-erased artists. The LPIPS divergence indicates preserving at least 500 artists is essential for retaining model performance. Additional results and analysis are provided in the Appendix.

D.2. Debiasing

Table D.1 shows the performance of the unified model compared to our individual debias models. On an average we find that unified models show a similar performance to debiased models.

Table D.2 displays the debiased results for all 36 individual professions from the WinoBias dataset. Our method consistently reduced bias and increased gender diversity in

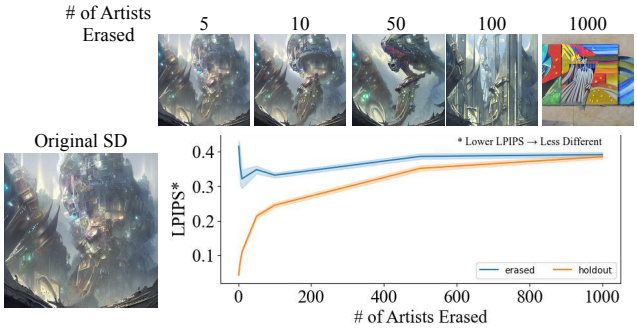

Figure D.1. The samples demonstrate model performance on erased artists after editing. We observed that erasing over 100 artists begins negatively impacting output for holdout artists, as evidenced by the increasing LPIPS after 100 erasures. Erasing 50 or fewer artists resulted in negligible interference on non-erased concepts.

Profession	SD	TIME + Preserve	Ours Debiased	Ours Unified
Assistant	0.19	0.57	0.14	**0.09**
Cook	0.82	0.15	**0.03**	0.14
Worker	1.00	0.15	**0.06**	0.09
Analyst	0.58	0.13	0.20	**0.03**
Doctor	0.78	0.41	0.20	**0.09**
WinoBias	0.67	0.31	**0.22**	0.27

Table D.1. Quantitative evaluation of profession debiasing for the unified model compared to an individual debiasing model. The metric Δ measures percentage deviation from equal gender ratios (Δ=0 denotes perfect equality). On average, the unified model achieves comparable debiasing performance to the individually finetuned model.

Stable Diffusion outputs for most professions. Additional qualitative results demonstrating gender debiasing can be seen in Figures D.9 and D.10. Figure D.11 provides further examples of improved racial diversity using our technique.

The algorithm outlined in the main paper describes our approach for debiasing diffusion model concepts by iteratively editing cross-attention weights. It takes as input the concepts to edit c_i, concepts to preserve c_j, and attribute text prompts to debias a_p. In a loop, current attribute ratio distributions R_{curr} are calculated for each concept using validation prompts and CLIP classification. R_{curr} is an $m \times p$ matrix, where m is the current edit list size and p the attribute count. The debiasing constants α_p are then computed proportionally to the difference between current and desired ratios, scaled by learning rate η. Once a concept is sufficiently debiased (within 5% of target), it is removed from the edit list and added to the preservation list.

This is done for 3 reasons - first, different concepts require varying levels of editing to remove bias, so they do not all debias at the same time. Once a concept is sufficiently debiased, we remove it from the edit list to avoid unnecessary

Figure D.2. Our method erases nudity content from pre-trained SD and has an advantage of erasing multiple concepts in I2P prompts. The figure shows percentage reduction in nudity classified samples for each body part type on I2P prompts compared to SD. "`Nudity`" erased model performs very similar to `ESD-x-1` as both the methods edit only cross attentions. Although, as noted in main paper, we find that our method results in a more finer edit and has better alignment with COCO.

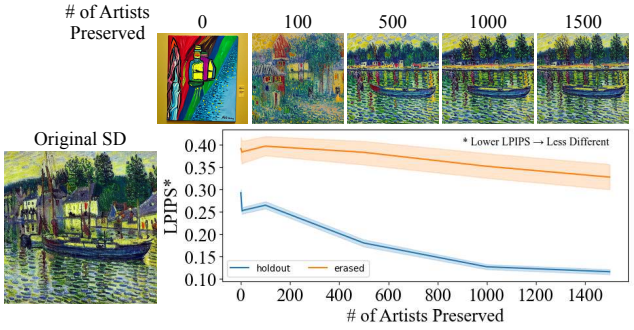

Figure D.3. The samples show the performance of edited models on holdout artist. We observe that preserving more artists is beneficial for reducing model's interference with nearby concepts. The plot shows LPIPS between original SD and models erasing 10 random artists with variable number of preservation artists. We find that preserving 500 artists and more has close to no interference on other surrounding concepts when erasing 10 artists.

generation of validation images for that concept. Second, editing one concept creates interference that can disrupt other concepts. Adding the debiased concept to the preservation list protects it from being affected by future edits. Third, careful asymmetric calculation of the debiasing constants α_p is required, unlike the symmetric constants used for erasing and moderating concepts. The optimal α_p values differ across concepts and attributes, necessitating the iterative tuning process.

D.3. Moderating NSFW

Figure D.2 displays the detailed erasure effects on different nudity classes classified by Nudenet. Our method demonstrates similar erasure to ESD-x for individual classes, while showing less interference on other concepts. The major advantage of our technique emerges in multi-concept erasure for I2P prompts and overall NSFW moderation, where our approach erases better than ESD methods across different NSFW classes.

D.4. Erasing Objects

Figures D.12- D.14 demonstrate effective object erasure using our method. One limitation of ESD-u was only partial removal of objects like churches, where major attributes such as crosses and tinted windows were erased but the building remained. In contrast, our approach shows stronger editing that clearly erases the full object.

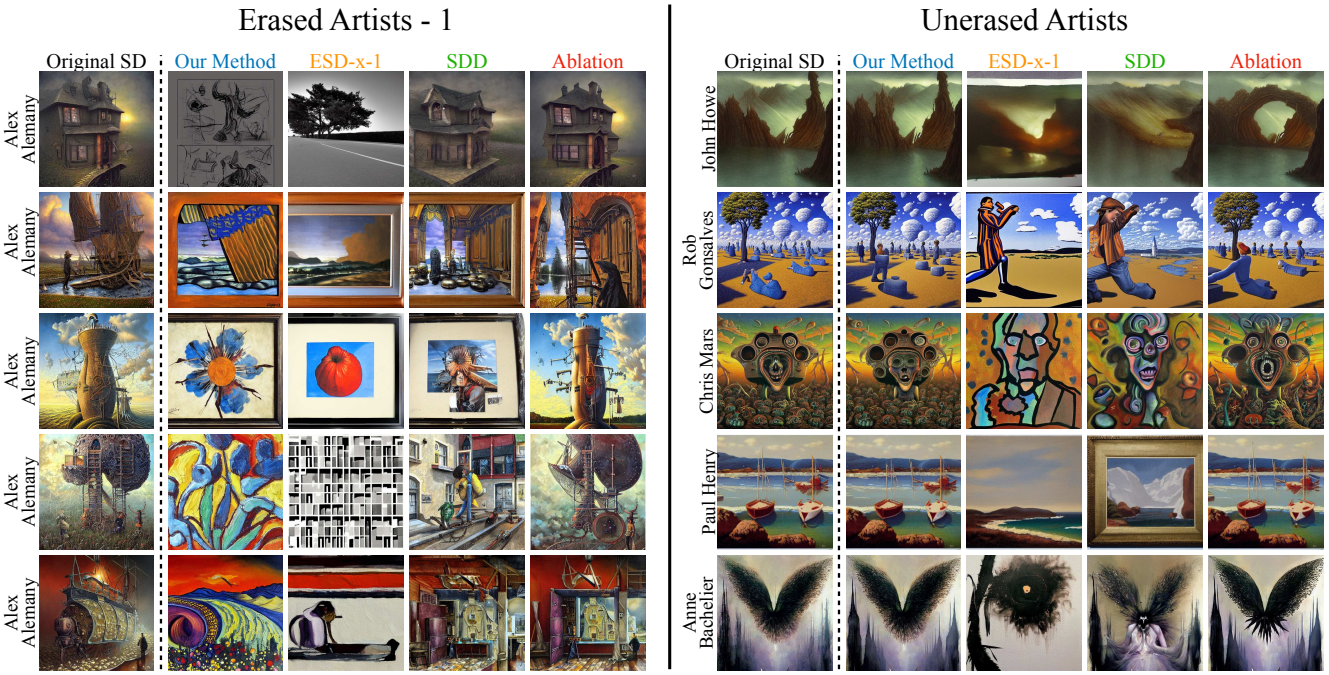

Figure D.4. Our method demonstrates a complete erasure of the intended artistic style and the least interference with the holdout artists that were neither erased nor preserved.

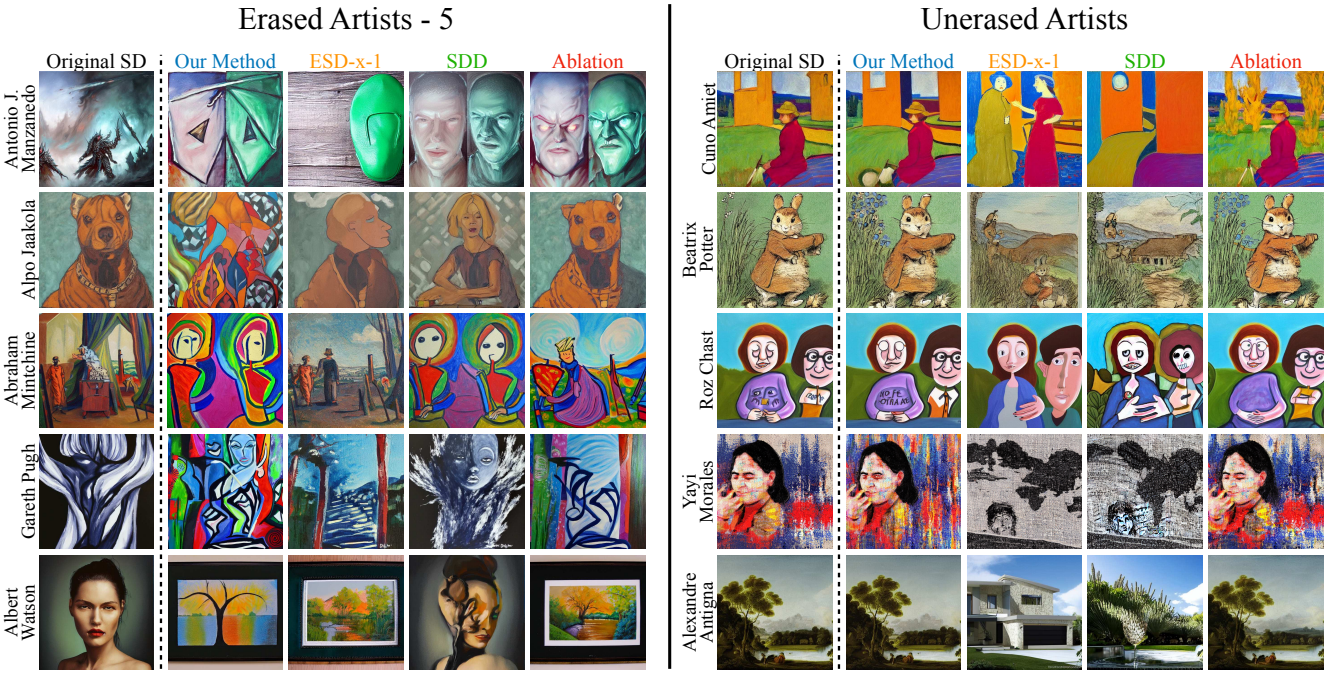

Figure D.5. Our method demonstrates strong multi concept erasure of intended artistic styles and the least interference with the holdout artists that were neither erased nor preserved.

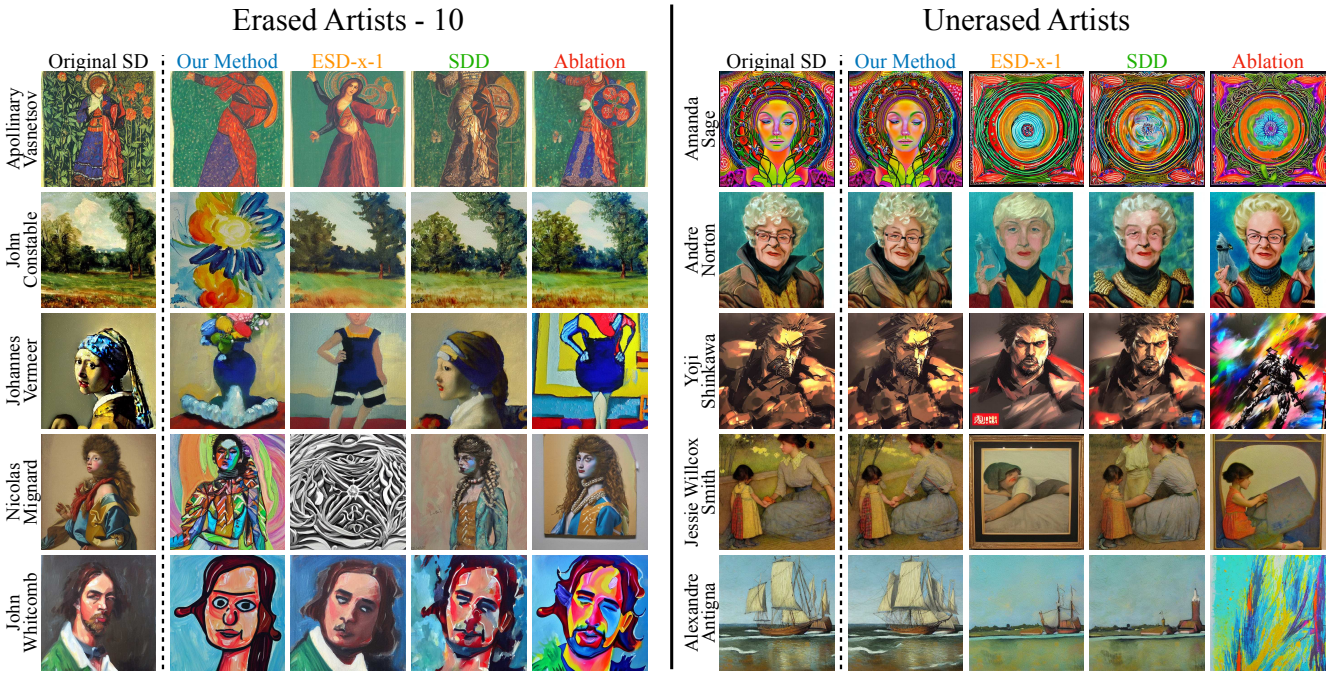

Figure D.6. Our method demonstrates strong multi-concept erasure of intended artistic styles and the least interference with the holdout artists that were neither erased nor preserved. Previous methods start showing interference effects when erasing 10 artists

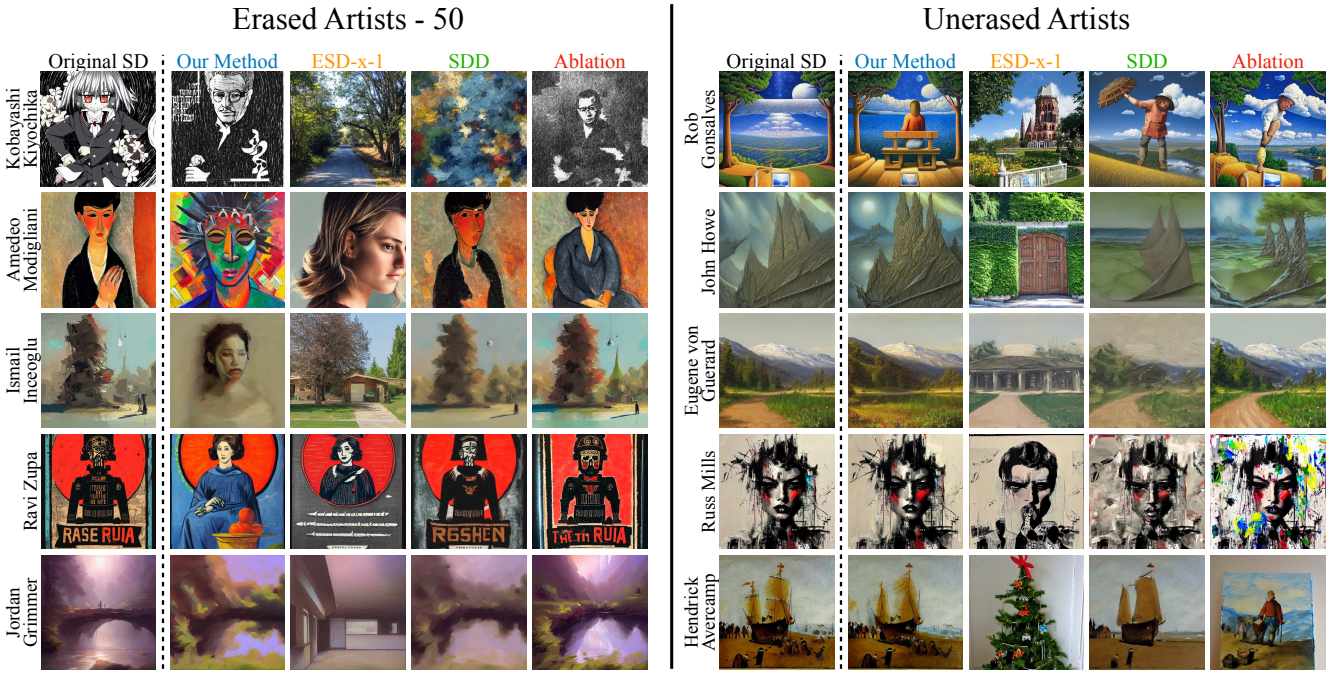

Figure D.7. Our method demonstrates strong multi-concept erasure of intended artistic styles and the least interference with the holdout artists that were neither erased nor preserved. Previous methods start showing interference effects when erasing 50 artists

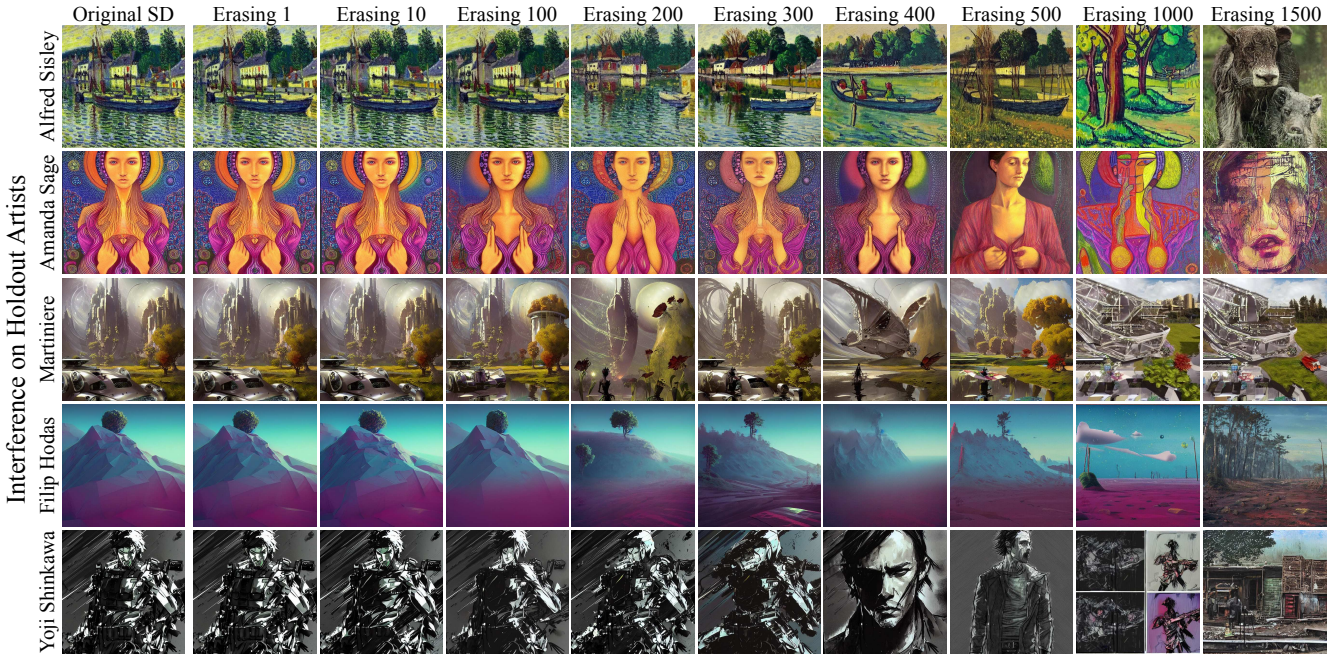

Figure D.8. The samples demonstrate edited model performance on holdout artists. We observed changes in output quality for holdout styles after erasing 300 artists. At 1000 erasures, the network starts to lose the artistic nuance in its generated images.

Figure D.9. Our method improves the gender representation of professions in the stable diffusion generated images. We find that the images precisely change the gender while keeping the rest of the scene intact.

Profession	Original-SD	Concept Algebra	Debias-VL	TIME	TIME + Preserve	Ours
Attendant	0.13 ± 0.06	0.23 ± 0.08	0.30 ± 0.04	0.50 ± 0.01	0.38 ± 0.11	0.09 ± 0.04
Cashier	0.67 ± 0.04	0.71 ± 0.10	0.23 ± 0.07	0.46 ± 0.01	0.23 ± 0.15	0.16 ± 0.06
Teacher	0.42 ± 0.01	0.46 ± 0.00	0.11 ± 0.05	0.34 ± 0.06	0.07 ± 0.06	0.06 ± 0.02
Nurse	0.99 ± 0.01	0.91 ± 0.05	0.87 ± 0.01	0.34 ± 0.03	0.30 ± 0.07	0.39 ± 0.07
Assistant	0.19 ± 0.05	0.20 ± 0.07	0.35 ± 0.15	0.32 ± 0.06	0.57 ± 0.08	0.14 ± 0.06
Secretary	0.88 ± 0.01	0.65 ± 0.07	0.65 ± 0.01	0.58 ± 0.09	0.71 ± 0.02	0.10 ± 0.10
Cleaner	0.38 ± 0.04	0.11 ± 0.06	0.18 ± 0.04	0.58 ± 0.07	0.79 ± 0.04	0.33 ± 0.07
Receptionist	0.99 ± 0.01	0.90 ± 0.08	0.74 ± 0.04	0.36 ± 0.10	0.24 ± 0.12	0.38 ± 0.01
Clerk	0.10 ± 0.07	0.11 ± 0.08	0.10 ± 0.04	0.40 ± 0.03	0.76 ± 0.05	0.23 ± 0.06
Counselor	0.06 ± 0.05	0.30 ± 0.03	0.10 ± 0.07	0.74 ± 0.08	0.41 ± 0.06	0.40 ± 0.02
Designer	0.23 ± 0.05	0.25 ± 0.12	0.48 ± 0.06	0.44 ± 0.06	0.23 ± 0.16	0.07 ± 0.05
Hairdresser	0.74 ± 0.11	0.37 ± 0.16	0.61 ± 0.04	0.32 ± 0.01	0.41 ± 0.09	0.16 ± 0.04
Writer	0.15 ± 0.03	0.07 ± 0.03	0.45 ± 0.04	0.54 ± 0.08	0.52 ± 0.08	0.31 ± 0.08
Housekeeper	0.93 ± 0.04	0.68 ± 0.18	0.80 ± 0.07	0.32 ± 0.03	0.68 ± 0.07	0.41 ± 0.05
Baker	0.81 ± 0.01	0.19 ± 0.04	0.72 ± 0.05	0.40 ± 0.04	0.19 ± 0.14	0.29 ± 0.08
Librarian	0.86 ± 0.06	0.66 ± 0.07	0.34 ± 0.06	0.26 ± 0.05	0.35 ± 0.01	0.07 ± 0.07
Tailor	0.30 ± 0.01	0.21 ± 0.05	0.33 ± 0.11	0.50 ± 0.03	0.03 ± 0.00	0.27 ± 0.01
Driver	0.97 ± 0.02	0.20 ± 0.07	0.65 ± 0.04	0.48 ± 0.09	0.17 ± 0.08	0.21 ± 0.07
Supervisor	0.50 ± 0.01	0.07 ± 0.03	0.43 ± 0.04	0.50 ± 0.07	0.42 ± 0.03	0.26 ± 0.04
Janitor	0.91 ± 0.05	0.71 ± 0.06	0.75 ± 0.05	0.36 ± 0.08	0.47 ± 0.12	0.16 ± 0.04
Cook	0.82 ± 0.04	0.48 ± 0.16	0.52 ± 0.07	0.38 ± 0.03	0.15 ± 0.10	0.03 ± 0.02
Laborer	0.99 ± 0.01	0.81 ± 0.06	0.98 ± 0.03	0.48 ± 0.08	0.24 ± 0.09	0.09 ± 0.02
Constr. worker	1.00 ± 0.00	0.95 ± 0.01	1.00 ± 0.00	0.40 ± 0.01	0.15 ± 0.05	0.06 ± 0.04
Developer	0.90 ± 0.03	0.74 ± 0.02	0.90 ± 0.04	0.50 ± 0.01	0.47 ± 0.07	0.51 ± 0.02
Carpenter	0.92 ± 0.05	0.84 ± 0.01	0.98 ± 0.01	0.52 ± 0.06	0.52 ± 0.05	0.06 ± 0.02
Manager	0.54 ± 0.06	0.15 ± 0.01	0.30 ± 0.05	0.38 ± 0.05	0.15 ± 0.01	0.19 ± 0.07
Lawyer	0.46 ± 0.08	0.13 ± 0.06	0.52 ± 0.05	0.64 ± 0.03	0.15 ± 0.03	0.30 ± 0.07
Farmer	0.97 ± 0.02	0.58 ± 0.09	0.97 ± 0.02	0.46 ± 0.02	0.27 ± 0.08	0.41 ± 0.01
Salesperson	0.60 ± 0.08	0.18 ± 0.05	0.07 ± 0.05	0.52 ± 0.05	0.05 ± 0.01	0.38 ± 0.05
Physician	0.62 ± 0.14	0.36 ± 0.10	0.70 ± 0.07	0.56 ± 0.06	0.49 ± 0.04	0.42 ± 0.01
Guard	0.86 ± 0.02	0.43 ± 0.12	0.48 ± 0.06	0.30 ± 0.10	0.10 ± 0.12	0.12 ± 0.07
Analyst	0.58 ± 0.12	0.24 ± 0.18	0.71 ± 0.02	0.52 ± 0.03	0.13 ± 0.05	0.20 ± 0.07
Mechanic	0.99 ± 0.01	0.65 ± 0.04	0.92 ± 0.01	0.38 ± 0.09	0.21 ± 0.04	0.23 ± 0.08
Sheriff	0.99 ± 0.01	0.38 ± 0.22	0.82 ± 0.08	0.22 ± 0.05	0.10 ± 0.05	0.10 ± 0.03
Ceo	0.87 ± 0.03	0.25 ± 0.11	0.37 ± 0.11	0.28 ± 0.04	0.18 ± 0.05	0.28 ± 0.03
Doctor	0.78 ± 0.04	0.40 ± 0.02	0.50 ± 0.04	0.58 ± 0.03	0.41 ± 0.08	0.20 ± 0.02
WinoBias	0.67 ± 0.01	0.43 ± 0.01	0.55 ± 0.01	0.44 ± 0.00	0.31 ± 0.00	0.22 ± 0.00

Table D.2. Our method has a consistent debiasing performance compared to previous inference and model editing methods. The presented metric Δ measures the percentage deviation from equal ratios ($\Delta = 0$ indicates perfect equal distribution across attributes) on 5 randomly picked professions out of 36 from the WinoBias dataset. On average, our method has the least deviation from the desired distribution.

Figure D.10. Our method improves the gender representation of professions in the stable diffusion generated images. We find that the images precisely change the gender while keeping the rest of the scene intact.

Figure D.11. Our method improves the racial diversity of professions in the pre-trained stable diffusion. We show images from the original SD and the corresponding images from the edited model for the same prompts and seeds for comparison. We find that our edited model has a better race representation.

Figure D.12. Our method demonstrates a complete erasure of the intended object and the least interference with unerased objects that are not explicitly preserved.

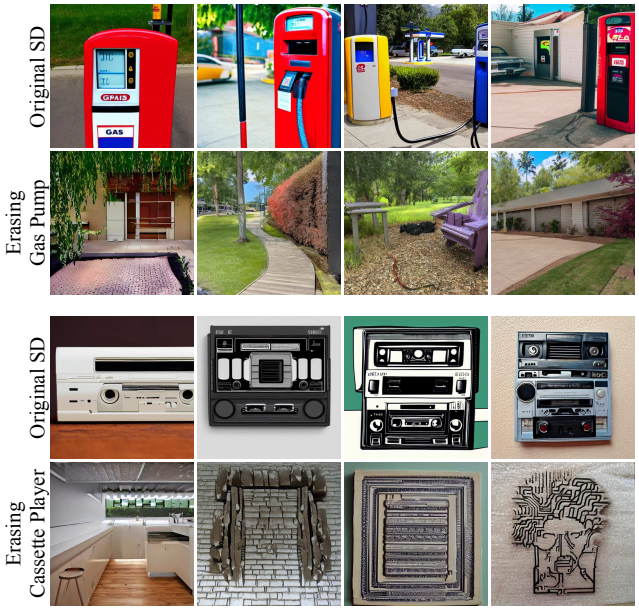

Figure D.13. Our method demonstrates a complete erasure of the intended object and the least interference with unerased objects that are not explicitly preserved.

Figure D.14. Our method demonstrates a complete erasure of the intended object and the least interference with unerased objects that are not explicitly preserved.