Roots Beneath the Cut: Uncovering the Risk of Concept Revival in Pruning-Based Unlearning for Diffusion Models

Supplementary Material

A. Importance of Signs vs. Magnitudes in Concept Revival

Under the ideal assumption that we have access to the weights of the pretrained model. We conduct experiments to analyze the importance of pretrained magnitudes and pretrained signs in reviving visual concepts. Specifically, we design three revival strategies: pretrained magnitudes with signs sampled from the distribution of the remained weight neuron, pretrained signs with sampled magnitudes via the same distribution and seed as the former strategy, and pretrained signs with our neuron max scaling. The results are shown in Table 4. The experimental results reveal a clear trend. When pretrained magnitudes are paired with sampled signs, the revived concepts exhibit even worse recognition accuracy than the unlearned model. In contrast, using pretrained signs together with sampled magnitudes yields a substantially stronger revival effect, demonstrating that correct signs are far more critical than correct magnitudes in reviving erased visual concepts.

B. Why we do Top-K Sign Retention

We observe that on the task of reviving erased concepts, the recovered weights across all visual concepts exhibit the same distribution pattern under our matrix completion method. Therefore, we select the recovered weights of *Golf Ball* as the representative example for analysis. We divide recovered weights into five equal-magnitude partitions, denoted as R_1–R_5 from largest to smallest, and similarly divide the pretrained weights into five groups, P_1–P_5, where a larger index corresponds to smaller magnitudes. To quantify their alignment, we compute, for each pair (R_i, P_i), the fraction of weight positions that overlap between the two groups, normalized by the number of R_i.

The resulting alignment matrix is reported in Table 5.

	P1	P2	P3	P4	P5
R1	**0.46**	0.27	0.16	0.07	0.01
R2	0.21	**0.26**	0.23	0.19	0.09
R3	0.14	0.19	**0.23**	0.24	0.19
R4	0.10	0.15	0.19	**0.24**	0.29
R5	0.09	0.13	0.17	0.22	**0.36**

Table 5. Alignment matrix between recovered groups (R_1–R_5) and pretrained groups (P_1–P_5).

We observe that each recovered magnitude group R_i exhibits a high positional overlap with its corresponding pretrained group P_i. **This indicates that our method effectively reconstructs the original magnitude structure of the pretrained model, placing large values back to positions that originally carried large weights.**

We further conduct a complementary analysis to examine how magnitude relates to sign correctness during recovery. Specifically, for each ratio $K \in \{0.2, 0.3, 0.5, 0.7, 0.9\}$, we select the top-$K$ recovered weights by magnitude as the **Top-K Magnitudes** group, and treat the remaining weights as the **Rest** group. By computing the sign accuracy within each group, we observe a clear trend: **weights with larger recovered magnitudes exhibit substantially higher sign correctness**.

The quantitative results are presented in Table 6.

K	Top-K Magnitudes' Signs Accuracy	Rest (1-K) Signs Accuracy
0.2	0.96	0.67
0.3	0.94	0.64
0.5	0.88	0.58
0.7	0.81	0.54
0.9	0.75	0.51

Table 6. Sign accuracy of recovered weights as a function of magnitude: Top-K groups consistently exhibit higher correctness than the Rest.

Combining these findings, matrix completion not only reconstructs the original magnitude structure of the pretrained model, but the recovered large-magnitude weights also exhibit significantly higher sign correctness. This makes **Top-K Sign Retention** a natural and necessary strategy for improving revival performance.

C. NSFW Content Revival

For the task of nudity revival, we evaluate our method on I2P [36], MMA [44] and Ring-A-Bell [39] datasets under three Top-K Sign Retention settings: $K = 0.4$, 0.6, and 0.8. The results are summarized in Table 7.

Table 4. Concept Erasure: Top-1 classification accuracy (%) of pre-trained, erased, and revived objects under ideal case using a pre-trained ResNet-50.

Classes	Pretrained [35] SD-V1-5	Concept [8] Prune ↓	Pretrained Weights Sampled Signs ↑	Pretrained Signs Sampled Weights ↑	Pretrained Weights Neuron-Max Scaling ↑
Church	86.6	8.2	8.0	68.0	**72.6**
Golf Ball	98.2	16.6	13.0	80.0	**96.4**
Gas Pump	82.8	6.8	2.0	59.0	**90.4**
Parachute	93.8	6.4	5.0	65.0	**48.8**
Chain Saw	70.8	0.00	0.4	25.0	**85.8**
French Horn	99.2	1.6	1.6	68.0	**89**
Mountain Bike	97.8	0.6	1.0	74.2	**90.2**
Starfish	100	32.2	16.8	91.4	**97.8**
Spider Web	96.8	22.2	13.2	73.6	**99.8**
School Bus	99	2.2	0.4	64.8	**96.8**
Racket	97.0	2.8	5.4	66.2	**89.4**
Candle	94.0	9.4	8.6	63.2	**67.6**
Average	93.0	9.1	6.3	66.5	**85.4**

Method	**I2P** (4703)	**MMA** (1000)	**Ring-A-Bell** (101)
Pretrained SD-v1.5	461	785	101
Concept Prune	74	57	22
Top-0.4	120	153	53
Top-0.6	78	173	46
Top-0.8	50	136	23

Table 7. Nudity revival performance across I2P, MMA and Ring-A-Bell datasets under different Top-K settings.

The three datasets I2P, MMA, and Ring-A-Bell contain 4703, 1000, and 101 prompts respectively. The number shown in the table corresponds to the number of images detected as containing NSFW content by the nudity detector [1].

We observe that after applying Concept Prune [8], the number of NSFW images generated by the model drops dramatically compared to the pretrained SD-v1.5. However, after applying our revival framework, the number of images detected as NSFW increases substantially across all three datasets.

This clearly demonstrates that our method is highly effective at reviving the erased NSFW concepts. The visualization of revival in NSFW contents are shown in Fig 8

D. Ablation on Top-K Sign Retention and Magnitude Strategies

We additionally perform ablation studies on the Parachute, Church, and Gas Pump revival tasks. The results are shown in Table 8, Table 9 and Table 10 respectively.

Table 8. Effect of Top-K and magnitudes on parachute revival.

Top-K Retention	**Neuron Sample**	**Neuron Average**	**NMS (Ours)**
Top 1.0	0.14	0.09	0.29
Top 0.8	0.16	0.13	0.37
Top 0.6	0.10	0.12	0.71
Top 0.4	0.10	0.11	0.59
Top 0.2	0.10	0.09	0.26

Table 9. Effect of Top-K and magnitudes on church revival.

Top-K Retention	**Neuron Sample**	**Neuron Average**	**NMS (Ours)**
Top 1.0	0.15	0.19	0.11
Top 0.8	0.18	0.20	0.12
Top 0.6	0.17	0.17	0.62
Top 0.4	0.13	0.16	0.55
Top 0.2	0.12	0.11	0.35

Table 10. Effect of Top-K and magnitudes on gas pump revival.

Top-K Retention	**Neuron Sample**	**Neuron Average**	**NMS (Ours)**
Top 1.0	0.12	0.13	0.01
Top 0.8	0.16	0.15	0.05
Top 0.6	0.15	0.13	0.30
Top 0.4	0.10	0.11	0.37
Top 0.2	0.10	0.10	0.24

Under a certain Top-K Sign Retention setting, the retained signs already capture the critical activation patterns associated with the erased visual concept. The results consistently show that our **Neuron Max Scaling** strategy further significantly amplifies such activation patterns by assigning magnitudes that reinforce retained signs. This leads to substantially improved revival performance compared to the other magnitude strategies. In addition, as the value of K decreases, the revival performance exhibits a clear trend of rising first and then falling, indicating that shutting down too many activation channels deteriorates the model's ability to reproduce the erased concept.

E. Attack diffusion unlearning methods that are based on targeting concept related weights

To verify the generalization of effectiveness of sign-domain for concepts and our revival framework, we implement our revival framework on Scissorhands[41] and SalUn[12] which both rely on gradients to identify concept-critical weights and then fine-tune those weights. Following their method, we prune those concept-critical weights and apply our NMS attack. As shown in Table 11, our NMS attack can effectively recover the unlearned accuracy from 21% to 66% and 68%. And we conclude that once the positions of concept-related critical weights are aware, our method can successfully attack and recover the unlearned concept while preserving overall model utility.

Table 11. NMS Attack on Scissorhands and SalUn.

Classes	Scissorhands Unlearn ↓	Scissorhands NMS Attack ↑	Preserved Class Accuracy ↑	SalUn Unlearn ↓	SalUn NMS Attack ↑	Preserved Class Accuracy ↑
Church	0.25	0.68	0.92	0.24	0.66	0.92
Golf Ball	0.23	0.76	0.93	0.24	0.80	0.91
Gas Pump	0.29	0.71	0.93	0.31	0.74	0.90
Parachute	0.20	0.62	0.91	0.21	0.65	0.92
Chain Saw	0.13	0.48	0.95	0.13	0.50	0.95
French Horn	0.13	0.34	0.90	0.14	0.44	0.91
Mountain Bike	0.11	0.67	0.93	0.13	0.71	0.90
Starfish	0.37	0.92	0.91	0.35	0.91	0.93
Spider Web	0.26	0.63	0.92	0.25	0.68	0.92
School Bus	0.15	0.79	0.93	0.18	0.83	0.91
Racket	0.21	0.68	0.91	0.20	0.66	0.91
Candle	0.17	0.66	0.92	0.14	0.63	0.92
Average	0.21	0.66	0.92	0.21	0.68	0.92

Table 12. Extend defense on all object erasure tasks

Gaussian Obfuscation σ_M	**Extend Defense**						**NMS Attack**	
	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-6}	10^{-2}
Church	0.08	0.09	0.09	0.08	0.06	0.00	0.06	0.08
Golf Ball	0.16	0.17	0.18	0.18	0.15	0.00	0.08	0.15
Gas Pump	0.07	0.07	0.09	0.05	0.03	0.00	0.01	0.07
Parachute	0.06	0.07	0.06	0.05	0.04	0.00	0.08	0.13
Chain Saw	0.00	0.02	0.04	0.04	0.05	0.03	0.01	0.01
French Horn	0.02	0.02	0.04	0.03	0.01	0.00	0.03	0.08
Mountain Bike	0.01	0.03	0.05	0.04	0.02	0.00	0.01	0.02
Starfish	0.32	0.30	0.27	0.25	0.20	0.14	0.04	0.31
Spider Web	0.22	0.25	0.27	0.20	0.16	0.05	0.03	0.18
School Bus	0.02	0.04	0.04	0.05	0.03	0.00	0.01	0.02
Racket	0.04	0.06	0.08	0.04	0.04	0.00	0.30	0.06
Candle	0.01	0.01	0.04	0.04	0.03	0.00	0.06	0.18
Average	0.08	0.09	0.10	0.09	0.07	0.02	0.06	0.11

F. Extensive evaluation on defense

We extend our defense on all classes under different Gaussian Obfuscation. Moreover, we further apply the attack on the defended model. The results in the table 12 (Left) shows that our defense is effective on all classes under different Gaussian Obfuscation σ_M. The results in the table 12 (Right) shows that our defense provides good resistance.

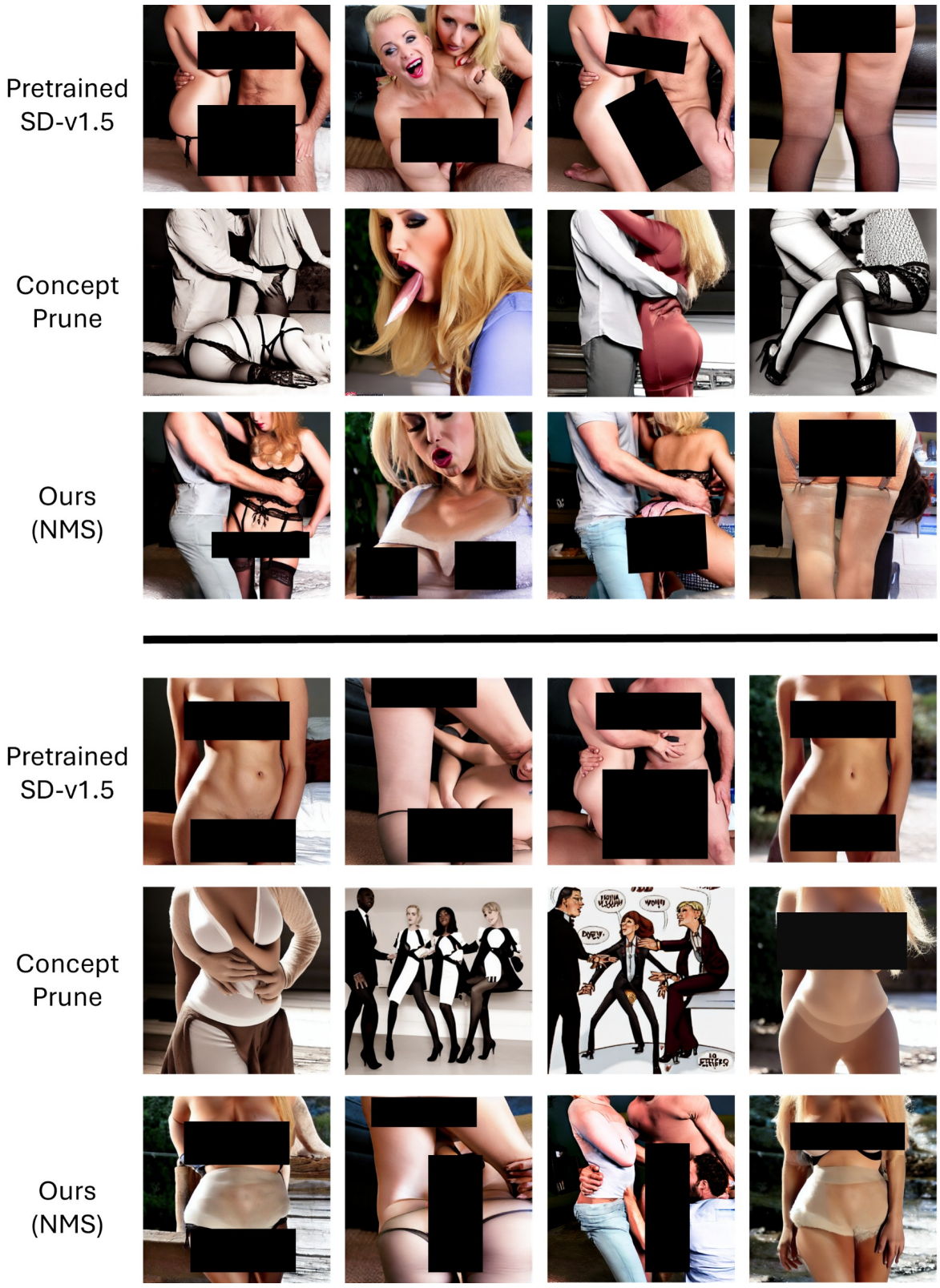
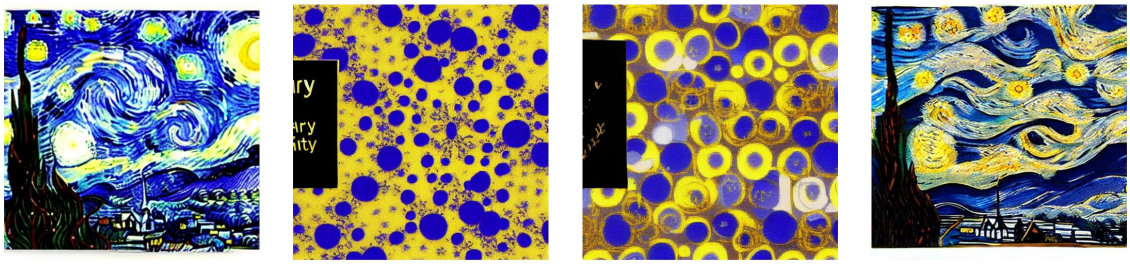


Figure 8. More visual results of revival on NSFW contents.

Starry Night by Vincent van Gogh



Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

The Church at Auvers by Vincent van Gogh

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

The Virgin and Child with St. Anne by Leonardo Da Vinci

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

Salvator Mundi by Leonardo Da Vinci

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

Figure 9. More visual results of revival on Van Gogh and Davinci.

Apparition of Face and Fruit Dish on a Beach by Salvador Dali

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

The Ecumenical Council by Salvador Dali

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

Women in the Garden by Claude Monet

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

Japanese Bridge by Claude Monet

Pretrained SD-v1.5 | Concept Prune | Quant Recover | Ours (NMS)

Figure 10. More visual results of revival on Dali and Monet.

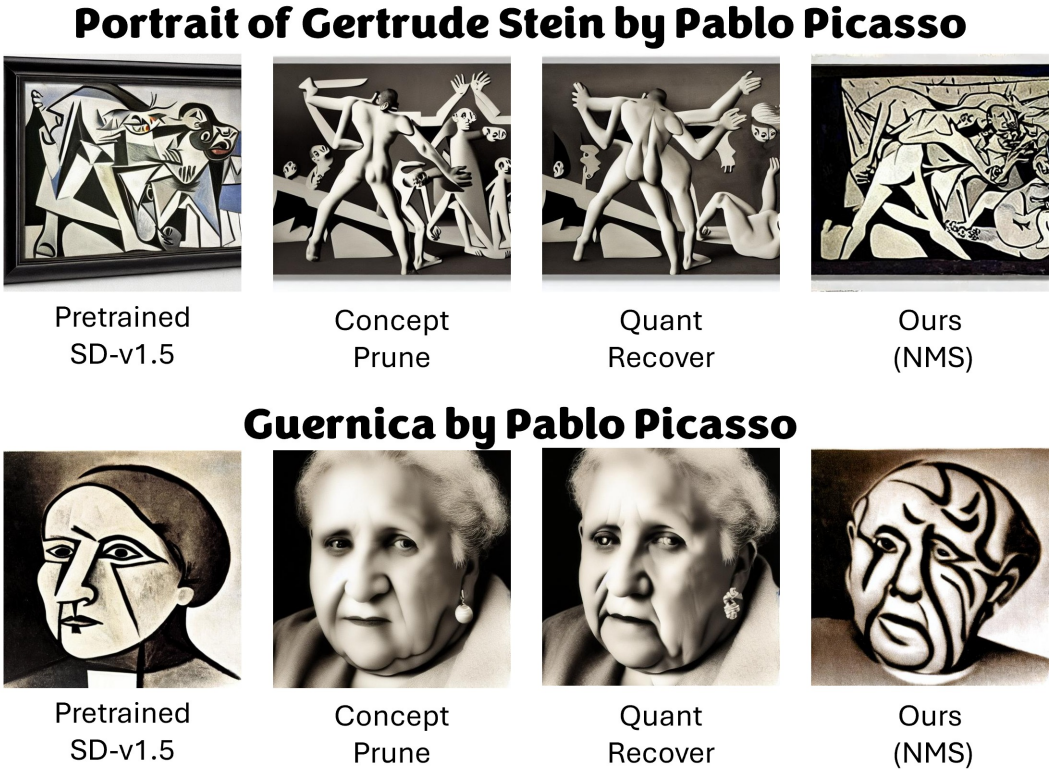


Figure 11. More visual results of revival on Picasso.

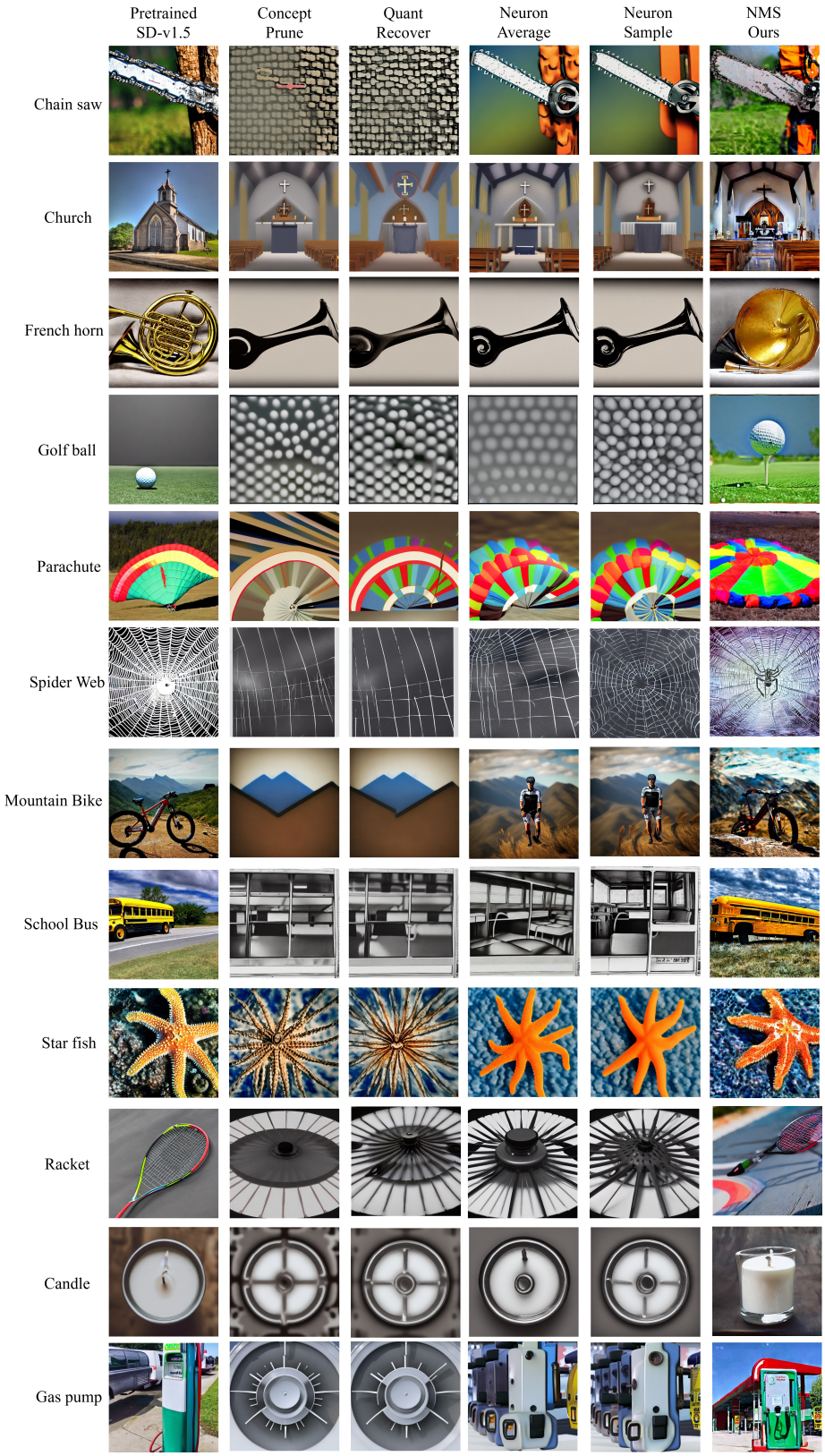


Figure 12. More visual results of revival on erased objects.