An Empirical Exploration of Continual Unlearning for Image Generation

Justin Lee [1] **Zheda Mai** [1] **Chongyu Fan** [2] **Wei-Lun Chao** [1]

Abstract

Machine unlearning typically assumes unlearning requests arrive simultaneously, whereas in practice, they often occur sequentially. We present the first systematic study of *continual unlearning* in text-to-image generation—after only a few requests, unlearned models based on popular methods drastically forget retained knowledge and produce degraded images. We attribute this behavior to cumulative parameter drift and explore add-on mechanisms that (1) mitigate drift and (2) remain compatible with existing unlearning methods. We show that constraining model updates and merging independently unlearned models are effective solutions, suggesting promising research directions. Taken together, our study positions continual unlearning as a fundamental problem in image generation, revealing open challenges to advance safe and accountable generative AI.

1. Introduction

Diffusion models (DMs) have demonstrated remarkable capabilities in text-to-image generation (Rombach et al., 2022; Kawar et al., 2023; Zhang et al., 2024a; Nichol et al., 2021), facilitating diverse applications (AdGen AI, 2025; Fotor, 2023). This versatility largely stems from massive internet-sourced training data, which inevitably introduces ethical and legal risks (Schramowski et al., 2023; TheStreet, 2023; Vinker et al., 2023). Recent regulations like CCPA (California Attorney General, 2018) grant users rights to request removals of their content from models, yet retraining DMs for each removal request is infeasible, requiring hundreds of thousands of GPU-hours (Gandikota et al., 2023).

Unlearning has emerged as a promising alternative to eliminate undesired generative capabilities (*e.g.*, an artistic style) without complete retraining (Hong et al., 2024; Gandikota et al., 2023; Kumari et al., 2023). Most existing methods assume unlearning requests arrive simultaneously (Wu et al., 2025; Gandikota et al., 2023; Kumari et al., 2023; Wu et al., 2024), while in practice, they often arrive sequentially[1].

To reflect this real-world scenario, we introduce the problem of ***Continual Unlearning (CU)*** for text-to-image generation, formalizing it as sequential removal of targeted generative capabilities. We present the first comprehensive study, accompanied by a new benchmark built upon UNLEARN-CANVAS (Zhang et al., 2024b) that considers two types of unlearning sequences—styles and objects (Figure 2).

Our findings reveal that popular unlearning methods—while effective at removing one or a few concepts simultaneously—suffer catastrophic failures in CU settings. After only a few requests, the model exhibits severe forgetting of retained knowledge, leading to significantly degraded image quality. Our analysis attributes this failure to *cumulative parameter drift*: successive unlearning steps push the model progressively farther from the pre-training manifold.

We explore several add-on mechanisms that can be seamlessly integrated into existing unlearning methods to mitigate parameter drift. These include: (1) regularizing updates relative to previously unlearned models; (2) merging independently unlearned models; (3) selectively updating the most critical parameters for unlearning the targeted concepts. Extensive results show the complementary effectiveness of these mechanisms regarding unlearning type (*i.e.*, styles or objects) and paired unlearning methods. Notably, updating as few as 0.1% of model parameters can already unlearn targeted concepts while preserving unrelated ones. These findings highlight the challenges and opportunities of continual unlearning and suggest promising directions for future research. Our **major contributions** are three-fold:

- We introduce and establish a benchmark for Continual Unlearning (CU) of text-to-image diffusion models.
- We find existing methods suffer critical deterioration and pinpoint cumulative parameter drift as the potential cause.
- We propose several effective add-on-style solutions, offering robust references for future research in CU.

[1]We include detailed related work in Appendix D.

[1]The Ohio State University [2]Michigan State University. Correspondence to: Justin Lee <lee.10369@buckeyemail.osu.edu>.

Published at MemFM Workshop at ICML 2025, Vancouver, Canada.

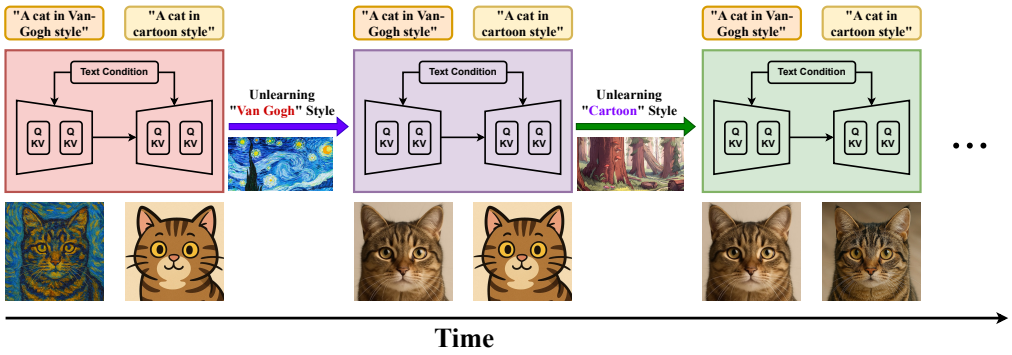

Figure 1. **Ideal outcomes of continual unlearning in image generation.** Initially, the model accurately generates images for prompts "A cat in Van Gogh (cartoon) style". After unlearning "Van Gogh," the model stops producing that style but maintains the cartoon style. Following the subsequent removal of "Cartoon," both style concepts are erased, while model retains the ability to generate "cat" images.

2. Preliminary

We investigate the unexplored questions in unlearning for generative models: *Do current techniques remain effective in continual settings? If not, how can we enhance them to efficiently handle sequential unlearning requests?*

We denote a generative model by $I = G_\theta(a)$, taking a text prompt a as input to guide content synthesis. Machine unlearning aims to modify θ into $\theta(q^\star)$ such that for prompts containing a target concept $q^\star \in a$, the generated image no longer reflects $q^\star$, as judged by a recognition model f_ϕ (i.e., $f_\phi(I) \neq q^\star$), while preserving all other concepts $q \neq q^\star$. More detailed preliminaries are in Appendix E.

3. Investigation of Unlearn Methods in Continual Setting

3.1. Problem Definition

In practice, a model may be asked to unlearn multiple concepts $Q = \{q_1, q_2, \cdots, q_N\}$. If all requests arrive together, one can apply simultaneous unlearning to obtain $\theta(Q)$; more commonly, requests arrive sequentially. Let θ_n denote the model after unlearning the first n concepts, defined recursively by $\theta_{n+1} = \theta_n(q_{n+1})$. After n steps, for any concept q that appears in the prompt (i.e., $q \in a$), the following should hold: if q has already been unlearned (i.e., $q \in \{q_1, \cdots, q_n\}$), then $f_\phi(I) \neq q$; otherwise (i.e., $q \notin \{q_1, \cdots, q_n\}$), we should have $f_\phi(I) = q$, where I is the image generated with prompt a.

We evaluate performance using Unlearning Accuracy (**UA**), In-Domain Retention Accuracy (**IRA**) and Cross-Domain Retention Accuracy (**CRA**). UA quantifies the proportion of cases where an unlearned concept $q^\star$ is no longer recognized, i.e., $f_\phi(I) \neq q^\star$. IRA/CRA measures the proportion of retained concepts $q \notin Q$ that remain correctly recognized, i.e., $f_\phi(I) = q$. IRA is for concepts semantically related to Q (e.g., other styles when unlearning a style), and CRA is for unrelated ones (e.g., objects when unlearning a style). For all of them, higher values mean better performance.

3.2. Setup

Data. We adopt the recently proposed UNLEARNCANVAS (Zhang et al., 2024b), which provides a fine-tuned Stable Diffusion model and high-accuracy classifiers across 60 styles and 20 objects. This framework standardizes assessment of unlearning accuracy and retention performance; see Appendix F.1 for dataset and evaluation details.

Unlearning Sequence. To evaluate continual unlearning under realistic conditions, we define two settings: sequential style and object unlearning, each involving the stepwise removal of 12 concepts. A shared held-out set of styles and objects enables consistent measurement of both in-domain and cross-domain retention; details in Appendix F.2.

Unlearning Methods. We evaluate two representative unlearning methods: Erased Stable Diffusion (ESD) (Gandikota et al., 2023), which uses reversed classifier-free guidance to suppress target concepts, and Concept Ablation (CA)(Kumari et al., 2023), which substitutes target concepts with anchor concepts. Details in Appendix F.3.

3.3. Existing Methods Fail To Unlearn Continually

While both ESD and CA effectively erase single concepts without significantly impacting retained knowledge, they face critical limitations in continual unlearning scenarios. Sequentially applying these methods—where each new unlearning step starts from the previously fine-tuned checkpoint—leads to catastrophic degradation of generation quality (Figure 2), as the model rapidly loses the ability to generate diverse, high-quality outputs (Figure 3.3).

Conversely, repeatedly restarting from the original checkpoint and simultaneously unlearning all accumulated concepts preserves performance but incurs prohibitive computational costs (Appendix B.2). This trade-off between computational efficiency and retention underscores the necessity of specialized continual unlearning methods that prevent

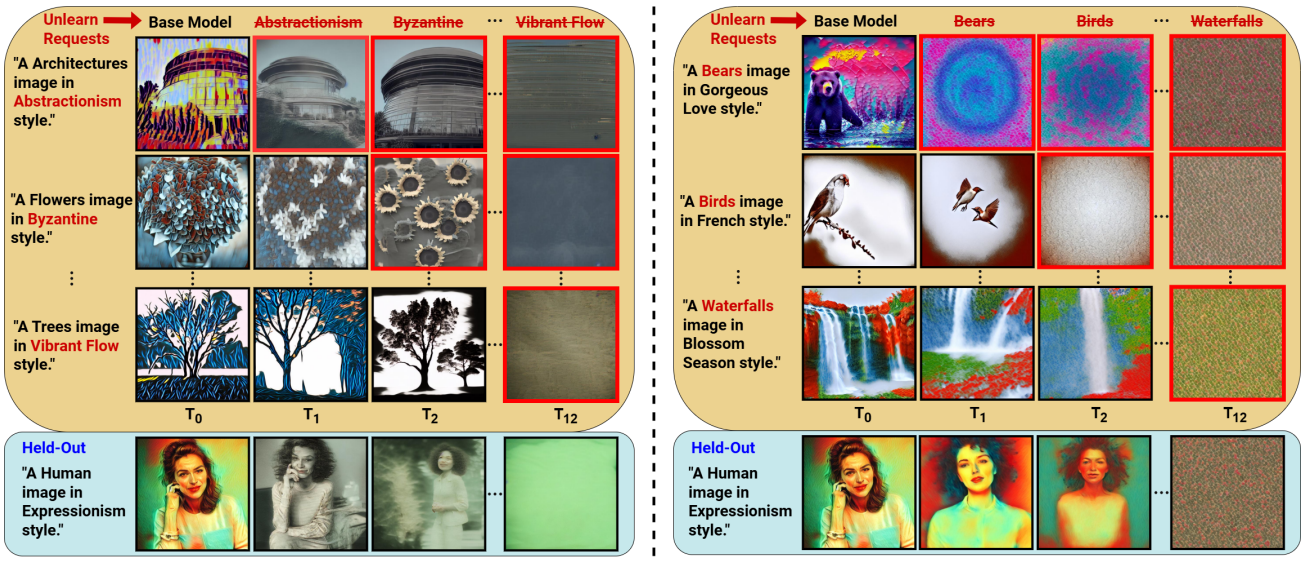

Figure 2. **Continual unlearning causes catastrophic degradation.** Left: sequentially unlearning artistic styles; right: sequentially unlearning objects. Starting from the base model (T_0), each step ($T_1 \cdots T_{12}$) removes a new concept (e.g., Abstractionism or Bears). Red boxes indicate images containing concepts already unlearned. Ideally, images without red boxes remain intact; however, continual unlearning progressively impairs the model's generation capabilities, culminating in the inability to produce meaningful images after multiple removals (final column).

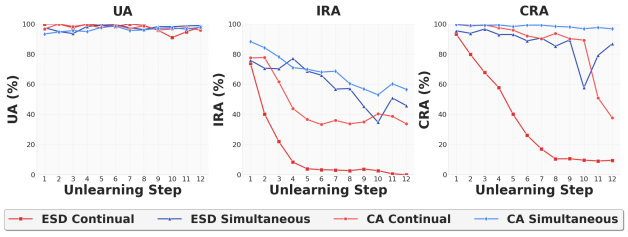

Figure 3. Continual Style Unlearning with CA and ESD rapidly degrade IRA and CRA, while Simultaneous shows improvement at the cost of repeated retraining.

catastrophic forgetting without frequent recomputation.

3.4. Why Do They Fail?

To investigate why continual unlearning leads to catastrophic degradation compared to simultaneous unlearning, we analyze the cumulative magnitude of parameter updates under each approach. Our analysis (Figure 4) reveals that sequential unlearning progressively induces greater parameter drift due to compounded fine-tuning steps, whereas simultaneous unlearning efficiently removes multiple concepts with significantly smaller cumulative updates. Contrary to the intuition that extensive parameter changes are necessary for multiple concept removals, our findings suggest that coordinated minimal interventions suffice. Thus, we propose **minimal intervention**—removing concepts with the smallest possible parameter adjustments—as a critical objective unique to continual unlearning, balancing effectiveness with computational efficiency.

4. Baseline Exploration for CU

Regularization. As cumulative parameter drift may drive CU failure, we propose two regularization strategies to limit update magnitudes, thus preserving the connection to the original parameter manifold: (i) L_1 regularization, which promotes sparsity by focusing updates into fewer parameters $\mathcal{L}_{L_1} = \mathcal{L}_{\text{unlearning}} + \lambda_1 \|\theta_n - \theta_{n-1}\|_1$; (ii) L_2 regularization, which encourages dispersed, smaller changes across many parameters $\mathcal{L}_{L_2} = \mathcal{L}_{\text{unlearning}} + \lambda_2 \|\theta_n - \theta_{n-1}\|_2^2$.

Model Merge. Observing that independently trained models incur lower update magnitude than continually fine-tuned models, we propose model merging as an alternative approach to CU. Unlike traditional methods that sequentially fine-tune from the previously unlearned model θ_{n-1}, we independently unlearn each concept directly from the base model θ_0 and then merge their weights. We adopt TIES-merging (Yadav et al., 2023) as our baseline, retaining only the top-k% largest parameter updates, effectively minimizing drift by limiting the number of parameters altered.

Selective Fine-tuning (SelFT). While L_1 regularization encourages sparse updates implicitly, optimization may still fail to pinpoint parameters critical for targeted concept removal. To address this, we utilize SelFT, explicitly identifying and constraining updates to the most crucial parameters. SelFT computes gradients of the unlearning loss using a single forward pass and selects the top $k\%$ parameters based on importance estimated via first-order Taylor expansion: $\text{Importance}(\theta_i) = |g_i \cdot \theta_i|$. Gradient updates are then restricted exclusively to these selected parameters using a

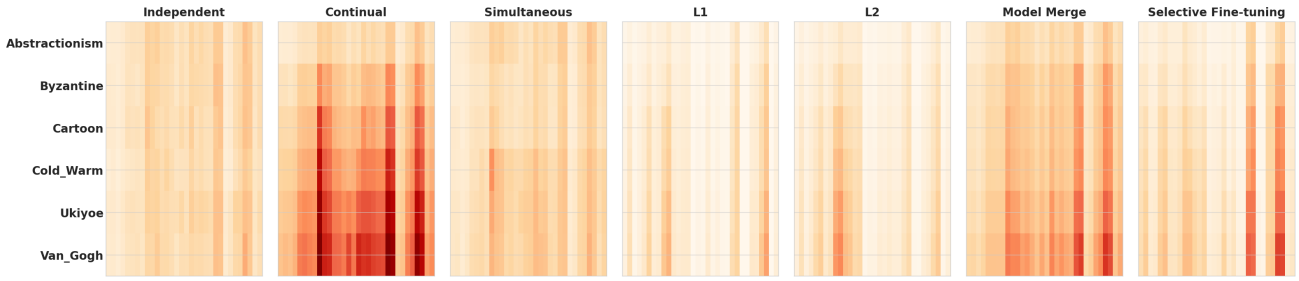

Figure 4. Parameter update magnitude: We visualize the L2 parameter distance from the base model across different unlearning approaches. The model unlearns the styles from top to bottom, continually using CA. While independent and simultaneous unlearning maintain minimal parameter drift (light colors), continual unlearning exhibits severe cumulative updates that progressively darken with each concept removal. Our proposed baselines effectively mitigate this drift: L1/L2 regularization constrains update magnitudes, model merging preserves the base model connection, and selective fine-tuning focuses on the most critical parameter for unlearning the targeted concepts.

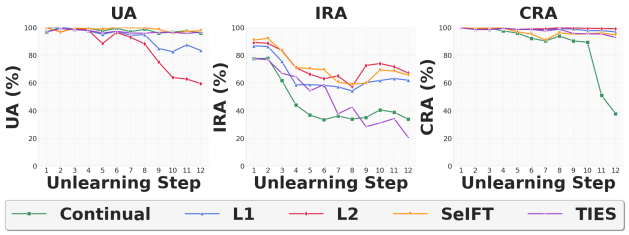

Figure 5. CA Object Unlearning, all proposed baselines outperform Continual in IRA and CRA, though L1 and L2 overwrites previous unlearning under long sequences. In this setting, SelFT achieves the best tradeoff between unlearning and retention.

binary mask.

4.1. Results

This empirical exploration contains numerous experimental details. Therefore, we distill several key findings below:

- **Overall:** Across all experimental settings, our proposed baselines consistently outperform naïve continual unlearning, yielding substantial gains in both IRA and CRA—even after extended unlearning sequences.
- **Compatibility Matter:** The optimal baseline varies significantly across different settings and algorithms. ESD Style performs best with TIES merging, CA Style with L1 regularization, ESD Object shows a three-way tie (excluding SelFT), and CA Object favors SelFT. The choice of both unlearning method and target concept type creates a complex optimization landscape, where the most effective minimal unlearning strategy must be tailored to the specific algorithm-domain pairing.
- **Unlearn Minimally:** We attribute the effectiveness of our baselines to their reduced cumulative parameter drift from the base model θ_0. As shown in Figure 4, each method introduces significantly smaller updates than standard continual unlearning, enabling stronger retention.
- **Unlearning is About Learning:** For remapping-based methods, the parameters updated during unlearning are influenced more by the anchor concept than by the target being removed. Figure 6 illustrates this effect, showing

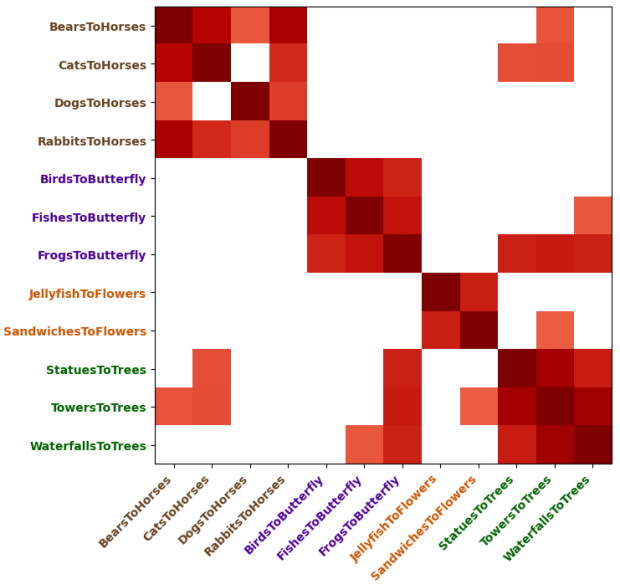

Figure 6. Weight update correlation matrix (top 30%) showing Pearson correlations between checkpoints from CA independent object unlearning. While unlearning is expected to align updates by target concept, we observe the opposite: updates cluster by anchor concept. This suggests that unlearning relies more on learning the anchor than erasing the target.

high similarity in update patterns when distinct target concepts are mapped to the same anchor.

5. Conclusion

We present the first systematic study of continual unlearning for image generation, reflecting real-world scenarios where unlearning requests arrive sequentially. Existing methods degrade quickly—forgetting retained concepts and generating low-quality images. We show that simple add-on mechanisms, such as regularization, merging, or selective updates, can significantly restore performance. These serve as strong baselines and highlight directions for further research.

References

AdGen AI. Finally. an ai ad generator that performs. publish ads in one click., 2025. URL https://www.adgenai.com/.

California Attorney General. California consumer privacy act (ccpa), 2018. URL https://oag.ca.gov/privacy/ccpa.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Davari, M. and Belilovsky, E. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pp. 270–287. Springer, 2024.

De Cao, N., Aziz, W., and Titov, I. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Fotor. Fotor: Online photo editor for everyone., 2023. URL https://www.fotor.com/.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.

Hong, S., Lee, J., and Woo, S. S. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21143–21151, 2024.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Javed, K. and White, M. Meta-learning representations for continual learning. *Advances in neural information processing systems*, 32, 2019.

Kadhe, S. R., Ahmed, F., Wei, D., Baracaldo, N., and Padhi, I. Split, unlearn, merge: Leveraging data attributes for more effective unlearning in llms. *arXiv preprint arXiv:2406.11780*, 2024.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6007–6017, 2023.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114:3521–3526, 2017.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Kuo, K., Setlur, A., Srinivas, K., Raghunathan, A., and Smith, V. Exact unlearning of finetuning data via model merging at scale. *arXiv preprint arXiv:2504.04626*, 2025.

Lee, N., Ajanthan, T., and Torr, P. H. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.

Mai, Z., Li, R., Kim, H., and Sanner, S. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3589–3599, 2021.

Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., and Sanner, S. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

Mallya, A., Davis, D., and Lazebnik, S. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 67–82, 2018.

Matena, M. S. and Raffel, C. A. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Mazumder, P., Singh, P., and Rai, P. Few-shot lifelong learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2337–2345, 2021.

Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

Nguyen, Q. H., Phan, H., and Doan, K. D. Unveiling concept attribution in diffusion models. *arXiv preprint arXiv:2412.02542*, 2024.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Ortiz-Jimenez, G., Favero, A., and Frossard, P. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Stoica, G., Bolya, D., Bjorner, J., Ramesh, P., Hearn, T., and Hoffman, J. Zipit! merging models from different tasks without training. *arXiv preprint arXiv:2305.03053*, 2023.

Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems*, 33: 6377–6389, 2020.

TheStreet. Users of midjourney text-to-image site claim issues with new update, 2023. URL https://www.thestreet.com/technology/users-of-midjourney-text-to-image-site-claim-issues-with-new-update.

Vinker, Y., Voynov, A., Cohen-Or, D., and Shamir, A. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42:1–13, 2023.

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Wu, J., Le, T., Hayat, M., and Harandi, M. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024.

Wu, Y., Zhou, S., Yang, M., Wang, L., Chang, H., Zhu, W., Hu, X., Zhou, X., and Yang, X. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504, 2025.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., and Liu, Z. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46:4115–4128, 2024a.

Zhang, Y., Fan, C., Zhang, Y., Yao, Y., Jia, J., Liu, J., Zhang, G., Liu, G., Kompella, R. R., Liu, X., et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.

Appendix

This appendix provides extended experimental results (Appendix A), implementation details (Appendix C), additional analysis supporting the main findings of the paper (Appendix B), detailed related work (Appendix D) and qualitative results (Appendix G). We include detailed comparisons of traditional and proposed unlearning baselines across artistic styles, objects, and celebrity concepts. Tables present metrics for unlearning accuracy (UA), in-domain retention accuracy (IRA), and cross-domain retention accuracy (CRA), highlighting the strengths and limitations of each approach under independent, continual, and simultaneous unlearning settings. Further, we offer insights into parameter update behavior and overlaps to motivate the design and interpretation of our proposed baselines.

A. Extended Experimental Results

A.1. Artistic Styles Unlearning

Traditional Baselines Table 1 demonstrates that both ESD and CA methodologies successfully unlearn individual style concepts when starting from the base model. Both approaches achieve perfect UA (100%) and maintain robust IRA scores, indicating effective single-concept unlearning without compromising model integrity.

However, as shown in Table 2, both methods demonstrate significant limitations in continual unlearning scenarios. While UA remains consistently high, IRA deteriorates dramatically to near-zero values for later styles, and CRA experiences substantial degradation, particularly for ESD. This indicates severe degradation of original model capabilities when sequentially unlearning multiple styles from previously unlearned checkpoints.

Table 3 illustrates that simultaneously unlearning all styles by restarting from the base model yields marked improvements in both IRA and CRA compared to continual unlearning. While this approach produces more balanced results, it introduces significant computational overhead, as it requires retraining from the base checkpoint for each unlearning request (Figure 12).

Proposed Baselines Table 4 examines the efficacy of L1 regularization during unlearning. This approach maintains strong UA while almost perfectly preserving CRA for CA (90-96%) and showing moderate CRA improvements for ESD. Both methods exhibit slight IRA improvements compared to the continual baseline. These results suggest L1 regularization effectively separates style erasure from object erasure by encouraging updates to consistent parameter regions for CA, while ESD benefits less dramatically. This difference may stem from CA mapping all styles to a generic "paintings" anchor, while ESD employs negative classifier-free guidance that potentially utilizes more disjoint parameter sets at each timestep.

Table 5 shows L2 regularization produces results comparable to L1 for ESD but inferior outcomes for CA. This pattern further supports the hypothesis that ESD updates parameters more broadly, as both sparsity-encouraging (L1) and dispersion-encouraging (L2) regularization yield similar results. Conversely, CA's performance decrease suggests unnecessarily dispersing parameter updates negatively impacts object preservation. This indicates L1 regularization may be particularly advantageous for erasure methods that naturally update similar regions across style concepts.

As demonstrated in Table 6, SelFT provides strong CRA performance for ESD through the first six styles, comparable to L1, before declining to match L1 performance by the twelfth style. This suggests selective parameter update sparsification initially isolates style from object erasure effectively, but after numerous updates, its benefits diminish as too many parameters become modified. CA exhibits a similar trend, with SelFT yielding strong IRA through the sixth style before gradually diminishing in effectiveness.

Table 7 reveals TIES merging as another powerful baseline for continual style unlearning. This approach produces substantial CRA improvements over standard continual unlearning for both ESD and CA, while also delivering the strongest IRA gains among all proposed methodologies. Notably, while TIES merging maintains computational efficiency comparable to continual unlearning, it introduces increased storage requirements as each unlearned checkpoint must be preserved and remerged for subsequent unlearning operations.

A.2. Objects Unlearning

Traditional Baselines Table 8 demonstrates the effectiveness of both ESD and CA methodologies in removing single object concepts, showing robust unlearning capabilities while maintaining strong retention metrics. Both approaches achieve high performance in isolation, establishing a solid baseline for object removal.

As evidenced in Table 9, both methodologies face significant challenges in continual object unlearning scenarios. Notably, ESD's IRA appears to stabilize around 12% while CRA continues to decline, suggesting some objects remain more deeply embedded within the diffusion model's parameters than artistic styles. CA exhibits the opposite pattern, maintaining stable CRA around 90% while IRA steadily declines, though CRA drops sharply around the eleventh unlearned object. This contrast suggests CA effectively isolates object concepts from style representations, but undergoes model collapse after numerous sequential updates away from the base model.

Table 10 illustrates the advantages of simultaneously unlearning all previously encountered objects from the pretrained model, demonstrating substantial improvements in both CRA and IRA metrics for both methodologies. However, a gradual performance degradation remains evident as the number of unlearned objects increases, indicating both approaches require progressively larger parameter updates as unlearning demands expand, potentially destabilizing the model's foundational capabilities.

Proposed Baselines Table 11 reveals an intriguing phenomenon with L1 regularization for CA. Through the first eight unlearning sequences (culminating with Rabbits), CA demonstrates significant IRA improvements ($\sim$20%) while maintaining near-perfect CRA and strong UA ($\sim$95%). However, beyond this threshold, UA begins fluctuating and declining as IRA stabilizes, suggesting extended sparsity-encouraged unlearning sequences eventually overwrite previous unlearning effects. Notably, this pattern does not manifest in CA's style erasure with L1, indicating objects may present greater erasure challenges than styles. ESD, meanwhile, maintains stable UA with more gradual IRA and CRA degradation compared to standard continual unlearning.

Table 12 demonstrates that L2 regularization exacerbates UA deterioration during extended unlearning sequences for CA, indicating dispersed parameter updates may be particularly vulnerable to unlearning reversal. ESD experiences slightly inferior IRA and CRA compared to L1, though these differences diminish over time, suggesting L1's sparsity constraints initially provide better concept isolation but offer diminishing returns after multiple updates.

As shown in Table 13, selective finetuning emerges as a particularly robust baseline for CA, delivering substantial IRA and CRA improvements without the UA rebounding observed with L1/L2 approaches. This suggests selective finetuning successfully identifies small, disjoint parameter sets for each target-to-anchor concept mapping, maintaining proximity to the base model while minimizing interference between unlearning updates. ESD initially demonstrates significant CRA and IRA gains for earlier sequences through Frogs, but these advantages rapidly diminish after extended unlearning, indicating that concentrated sparse updates eventually push the model too far from its pretrained state.

Table 14 confirms model merging as another competitive baseline for object erasure. CA's implementation of TIES merging successfully prevents the CRA collapse observed in continual unlearning. Interestingly, while continual unlearning sees IRA stabilize around 30-40%, TIES merging allows IRA to degrade below 30%. Through the first eight sequences, TIES outperforms continual unlearning in IRA, but subsequently declines while maintaining high CRA. This pattern suggests TIES merging effectively isolates object concepts from style representations but, by specifically targeting object-related parameters, accelerates deterioration in the model's ability to generate other objects. For ESD, the data reveals initially promising performance with substantial improvements to both IRA and CRA metrics through the first six unlearned objects. However, this is followed by a harsh drop in IRA, which ultimately stabilizes around 20%, while CRA continues its downward trajectory without stabilization. This persistent degradation in CRA suggests ESD may be less effective at maintaining clear parameter boundaries between object-specific and style-specific representations, resulting in more widespread model deterioration as unlearning sequences extend beyond initial objects.

A.3. Celebrities Unlearning

Traditional Baselines Table 15 demonstrates that both ESD and CA methodologies achieve reliable celebrity erasure when applied to individual concepts, successfully removing target celebrities from generation while maintaining robust retention accuracy of approximately 90%. This establishes that both approaches possess the fundamental capability to eliminate celebrity likenesses without significantly compromising the model's ability to generate other individuals.

However, as illustrated in Table 16, continual celebrity unlearning reveals severe performance degradation across sequential operations. When unlearning a sequence of six celebrities, retention accuracy plummets dramatically from around 90% to approximately 20%, indicating substantial deterioration in the model's capacity to generate non-target celebrities. This pattern suggests that celebrity representations may be more deeply intertwined within the diffusion model's parameter space than initially anticipated, making sequential erasure particularly challenging without affecting related identity features.

Table 17 confirms the computational trade-off inherent in simultaneous unlearning approaches. By consistently restarting from the base model to unlearn all previously encountered celebrities simultaneously, both methodologies achieve significant improvements in retention accuracy compared to continual approaches. However, this performance gain comes at considerable computational cost, as each unlearning request necessitates complete retraining from the original checkpoint, making this approach potentially prohibitive for practical deployment scenarios.

Proposed Baselines Table 18 reveals a particularly intriguing phenomenon with L1 regularization, where the sparsity constraint produces markedly different effects across the two methodologies. For ESD, L1 regularization appears to exacerbate performance degradation by inadvertently undoing previous unlearning operations. This deterioration likely stems from ESD's tendency to utilize similar parameter sets when unlearning different celebrities—the sparsity constraint forces repeated updates to identical parameter locations, leading to accumulated update magnitudes that progressively push the model further from its pretrained state while simultaneously overwriting previous unlearning effects. Conversely, CA demonstrates more resilient behavior under L1 constraints. While the sparsity encouragement somewhat diminishes its ability to deeply unlearn individual celebrities, it simultaneously improves celebrity isolation capabilities, resulting in enhanced retention accuracy that compensates for the reduced unlearning depth.

Table 19 shows that L2 regularization perpetuates similar challenges for ESD, with evidence suggesting continued preference for updating overlapping parameter sets across different celebrity targets. The progressive decline in unlearning accuracy coupled with deteriorating retention performance indicates that accumulated update magnitudes continue to destabilize the model's foundational capabilities. CA, however, exhibits improved performance compared to its L1 counterpart, demonstrating enhanced ability to thoroughly unlearn individual celebrities while achieving modest retention accuracy improvements through reduced update magnitudes relative to standard continual unlearning.

As demonstrated in Table 20, selective fine-tuning yields mixed results that further illuminate the fundamental differences between these unlearning approaches. ESD experiences notable decreases in both unlearning and retention accuracy, reinforcing the hypothesis that celebrity unlearning consistently targets similar parameter groups regardless of the specific individual being erased. This parameter overlap leads to compounding updates that simultaneously undo previous unlearning operations and push the model increasingly distant from its pretrained state. In stark contrast, CA demonstrates progressive improvements in retention accuracy while maintaining robust unlearning performance. This success suggests that selective parameter identification enables CA to locate disjoint parameter sets for each celebrity mapping, facilitating minimal yet effective updates that avoid interference with previous unlearning operations while maintaining proximity to the base model.

Table 21 reveals model merging as the most effective baseline among all proposed approaches for celebrity unlearning. Both ESD and CA achieve strong unlearning accuracy while realizing substantial retention accuracy improvements compared to traditional continual approaches. This superior performance can be attributed to TIES merging's sophisticated pruning mechanism applied to task vectors, which more effectively isolates the specific unlearning effects while preserving the merged model's proximity to the base checkpoint. The pruning step appears particularly beneficial for celebrity unlearning, where parameter interference between different individual representations poses significant challenges for sequential unlearning approaches.

B. Additional Analysis

B.1. Update Location

Importance of Anchor Concept For unlearning methods that re-map target concepts to anchor classes, the parameters that are updated depend largely on the anchor concept rather than the target concept being unlearned. This phenomenon is demonstrated in Figure 7, which shows task vectors (the difference in weights between fine-tuned and base models) for concept ablation (CA) independent object unlearning, where the update activations exhibit remarkable similarity for checkpoints that map to the same anchor concept. The correlation analysis in Figure 8 further supports this finding, revealing a clear block structure in the Pearson correlation coefficients between task vector pairs, indicating that unlearning different targets to the same anchor produces highly similar parameter updates. This surprising insight suggests that anchor-based unlearning methods function more as learning tasks than traditional removal approaches—rather than precisely targeting and updating parameters where the original concept resides, these methods essentially relearn the representation of the anchor concept, fundamentally reframing unlearning as a form of concept substitution rather than concept elimination.

For comparison, the same correlation matrix is shown for unlearning method ESD in Figure 9. Unlike CA's structured block

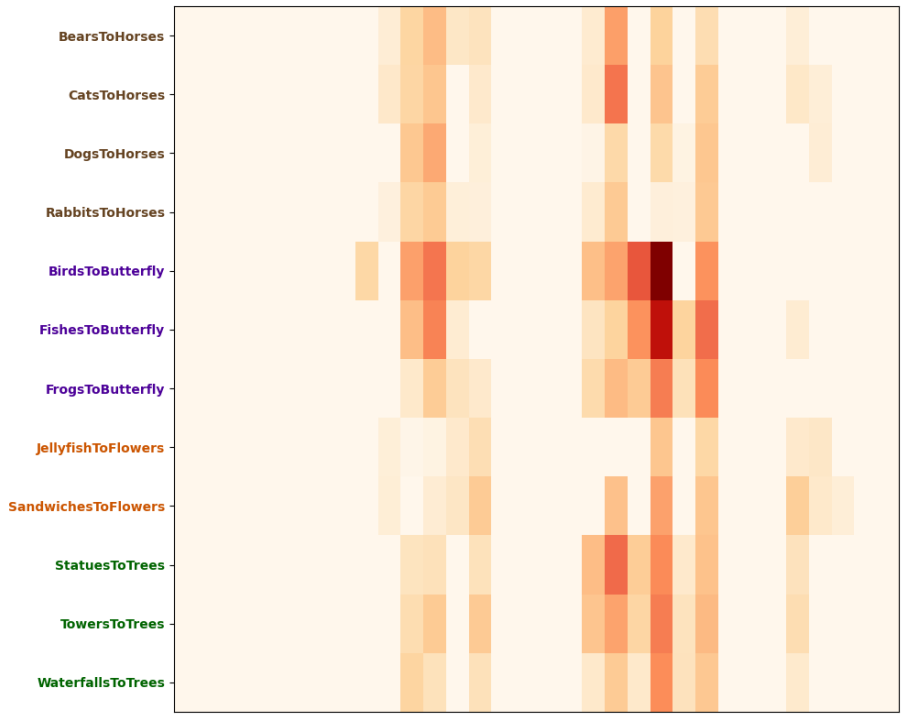

Figure 7. Task vector magnitude heatmap for unlearning method CA showing parameter modifications during object unlearning. Each row represents an independently trained checkpoint, with checkpoints grouped by anchor object class. Each column represents a trainable parameter. Darker colors indicate larger parameter changes from the base model. Only the top 30% of weight updates are displayed to highlight the most significant modifications.

pattern, ESD exhibits much more random correlation values across the matrix. This fundamental difference stems from ESD's approach: rather than explicitly mapping targets to predefined anchor concepts, ESD employs negative classifier-free guidance that implicitly determines concept substitutions during the unlearning process. This stochastic nature of anchor selection introduces greater variability in parameter updates, resulting in the more dispersed correlation pattern observed in the matrix.

Parameter Update Overlap The superior performance of selective fine-tuning in continual object unlearning, especially for CA (Table 13), can be attributed to its ability to reduce parameter update overlap. We analyze parameter overlap by computing task vectors as the difference between consecutive checkpoints in the unlearning sequence, revealing how parameter modifications accumulate over time. Standard continual unlearning exhibits high parameter overlap, with nearly all previously updated parameters being modified again in subsequent steps (Figure 10), causing individual parameters to drift significantly from their original values in the base model. In contrast, selective fine-tuning strategically identifies disjoint sets of parameters that are most critical for unlearning each specific concept. This parameter selection strategy substantially reduces overlap between unlearning steps (Figure 11), ensuring that the same parameters are not repeatedly updated throughout the sequence. By avoiding excessive modification of individual parameters, selective fine-tuning better preserves the original model's capabilities while achieving effective concept removal, explaining its superior retention performance in continual unlearning scenarios.

B.2. Simultaneous Training Costs

While simultaneous unlearning offers clear advantages in both IRA and CRA, it often incurs significantly higher computational costs. This is because each unlearning iteration begins from the pretrained model and must re-unlearn all previously seen concepts. As the number of concepts increases, so does the number of training steps typically required to reach full unlearning. In our experiments, we approximate full unlearning by halting training once frequent sample evaluation yields a UA of 99%. However, because each additional concept requires a longer training sequence, the total training cost grows

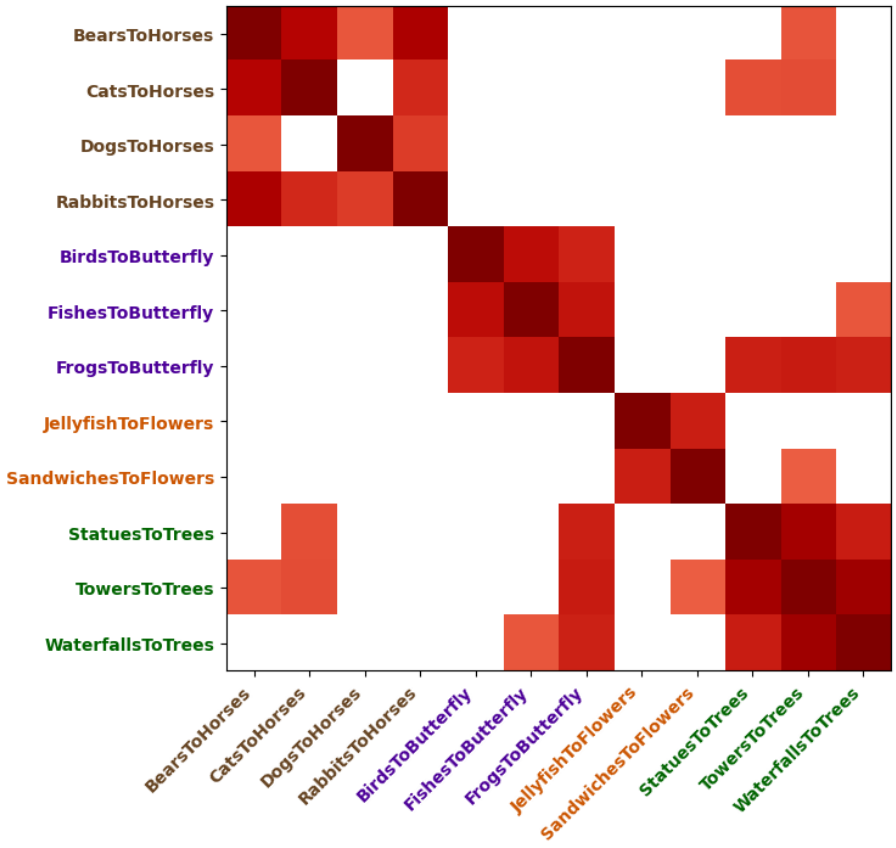

Figure 8. Task vector correlation matrix showing the top 30% of correlation values. Each cell represents the Pearson correlation coefficient between task vectors of checkpoint pairs from CA independent object unlearning. Checkpoint labels are colored by shared anchor concept, revealing clear block structure where same-anchor unlearning exhibit significantly higher correlations than cross-anchor unlearning. White cells indicate correlations below the 30% threshold.

non-linearly—unlike the fixed linear cost associated with continual unlearning. As unlearning sequences lengthen, the cumulative cost of simultaneous unlearning rapidly escalates, approaching exponential growth (Figure 12).

C. Implementation Details

This section provides comprehensive details on our evaluation methodology and implementation specifics for all unlearning approaches examined in this study. We describe the base models, evaluation protocols, and hyperparameter configurations to ensure reproducibility and clarity of our experimental framework.

C.1. Evaluation

Our evaluation framework builds upon established benchmarks for concept unlearning in diffusion models. All style and object unlearning algorithms were applied to the finetuned Stable Diffusion checkpoint from UnlearnCanvas, ensuring consistency in our experimental baseline. For evaluation of both style and object unlearning efficacy, we utilized the specialized classifier provided by UnlearnCanvas, which enables quantitative assessment of concept presence and retention.

For celebrity unlearning experiments, we diverged slightly by utilizing the pretrained base Stable Diffusion model available on Huggingface as our starting point. The evaluation of these experiments employed the GIPHY celebrity classifier, using classifier error rates to measure Unlearning Accuracy (UA) and classifier accuracy on held-out celebrities to determine Retention Accuracy (RA). To facilitate clear interpretation, we normalized RA values by dividing by the average pre-unlearning classifier accuracy of approximately 95.75%, thus providing a relative measure of concept retention.

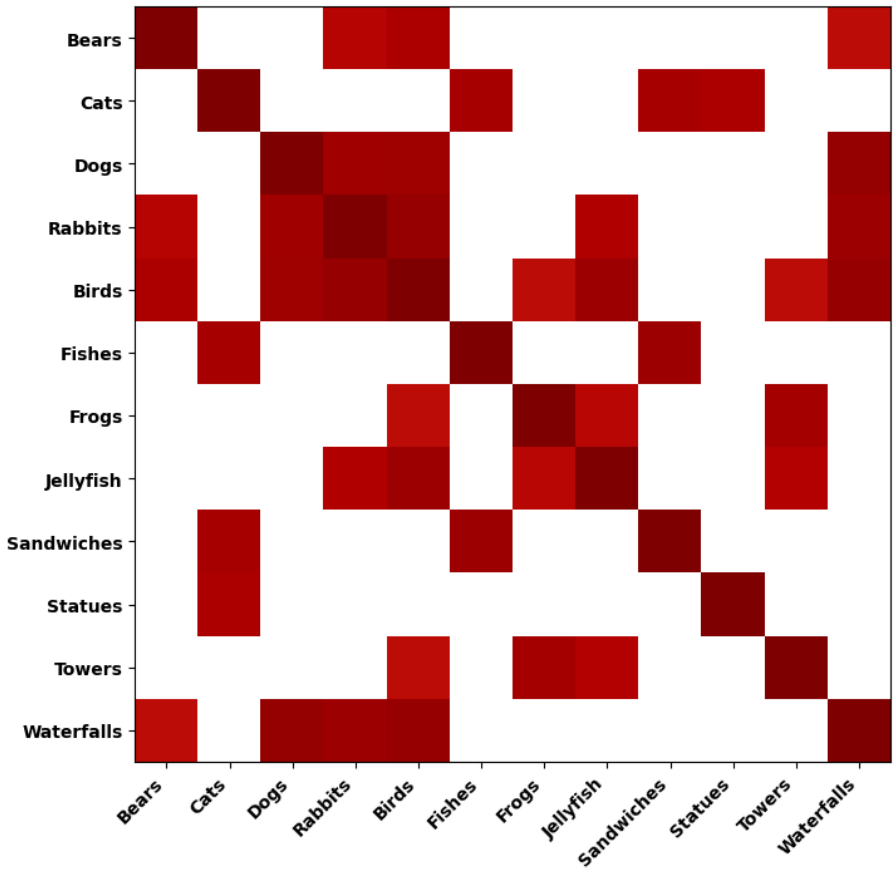

Figure 9. Task vector correlation matrix for ESD independent object unlearning, showing the top 30% of Pearson correlation coefficients between checkpoint pairs. The random dispersion of correlation values demonstrates ESD's varied parameter updates across different concepts, which contrasts with CA's more structured approach using anchor concept mapping.

C.2. Independent

For the independent unlearning scenario, we adapted implementation parameters based on concept type and algorithm requirements. ESD Style experiments followed the original paper's recommended hyperparameters with a learning rate of 1e-5, 1000 training steps, and updates restricted to cross-attention parameters. For ESD Object, we maintained the 1000-step protocol but scaled down the learning rate from the default 1e-5 to 5e-6, as our preliminary experiments indicated that the higher learning rate caused severe IRA and CRA degradation even when unlearning a single object concept. This modification enabled successful unlearning while ensuring realistic assessment of performance degradation in sequential scenarios.

CA Style implementation adhered to the original paper's guidelines, using a base learning rate of 2e-6 and restricting updates to key and value parameters in cross-attention layers. However, we extended the training duration from the originally recommended 110 steps to 1000 steps after observing insufficient unlearning performance (50% UA) with the shorter training schedule. Following the established CA methodology, each target style was mapped to the generic concept of "paintings," with an anchor dataset constructed from 200 LAION-sourced painting prompts used to generate images with the UnlearnCanvas checkpoint.

For CA Object, we maintained the base learning rate of 2e-6 and cross-attention KV parameter update restrictions, but further extended training to 4000 steps to achieve complete concept unlearning (100% UA). Each target object was mapped to a semantically related object from the held-out retention set, ensuring fair evaluation given the classifier's recognition constraints. Specifically, we mapped Bears, Cats, Dogs, and Rabbits to Horses; Birds, Fishes, and Frogs to Butterfly; Jellyfish and Sandwiches to Flowers; and Statues, Towers, and Waterfalls to Trees. This mapping strategy follows the CA

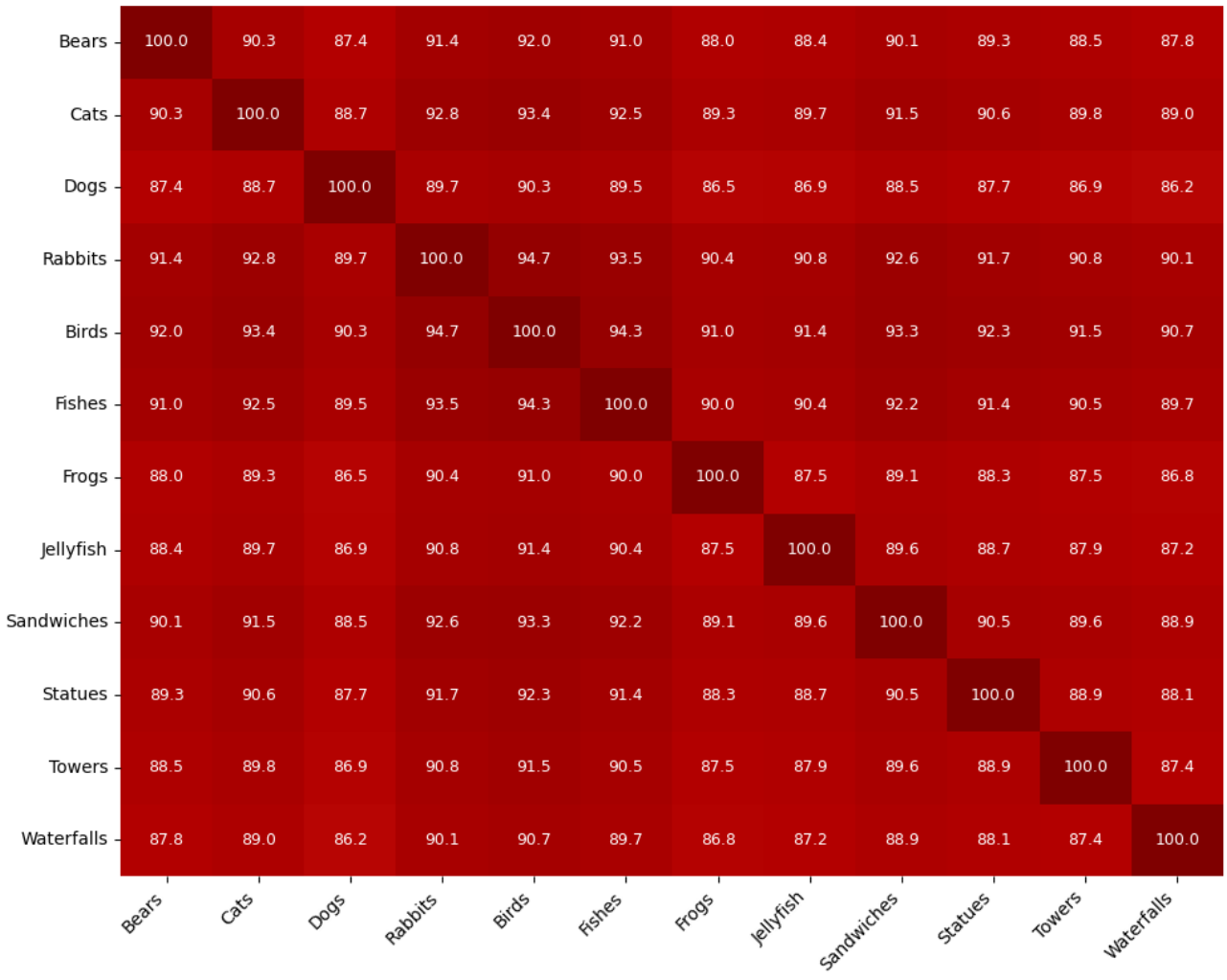

Figure 10. Parameter overlap matrix for CA continual object unlearning showing incremental weight changes between consecutive unlearning steps. High overlap values indicate that previously modified parameters are repeatedly updated, leading to cumulative parameter drift from the base model.

methodology of associating targets with visually similar or parent concepts.

ESD Celebrity experiments utilized identical parameters to ESD Style (learning rate of 1e-5, 1000 steps, cross-attention parameter updates), while CA Celebrity followed our CA Object configuration. For CA Celebrity, each target individual was mapped to the generic concept of "man" or "woman" based on gender, with prompts generated by GPT-3.5 Turbo and images produced by the base Stable Diffusion model.

C.3. Continual

Our continual unlearning implementation maintained consistent hyperparameters with the independent setting for all methods. The key distinction is that each unlearning operation initiated from the previously unlearned model checkpoint rather than the original base model. This sequential approach enables us to evaluate how performance degrades across successive unlearning operations and assess catastrophic forgetting effects.

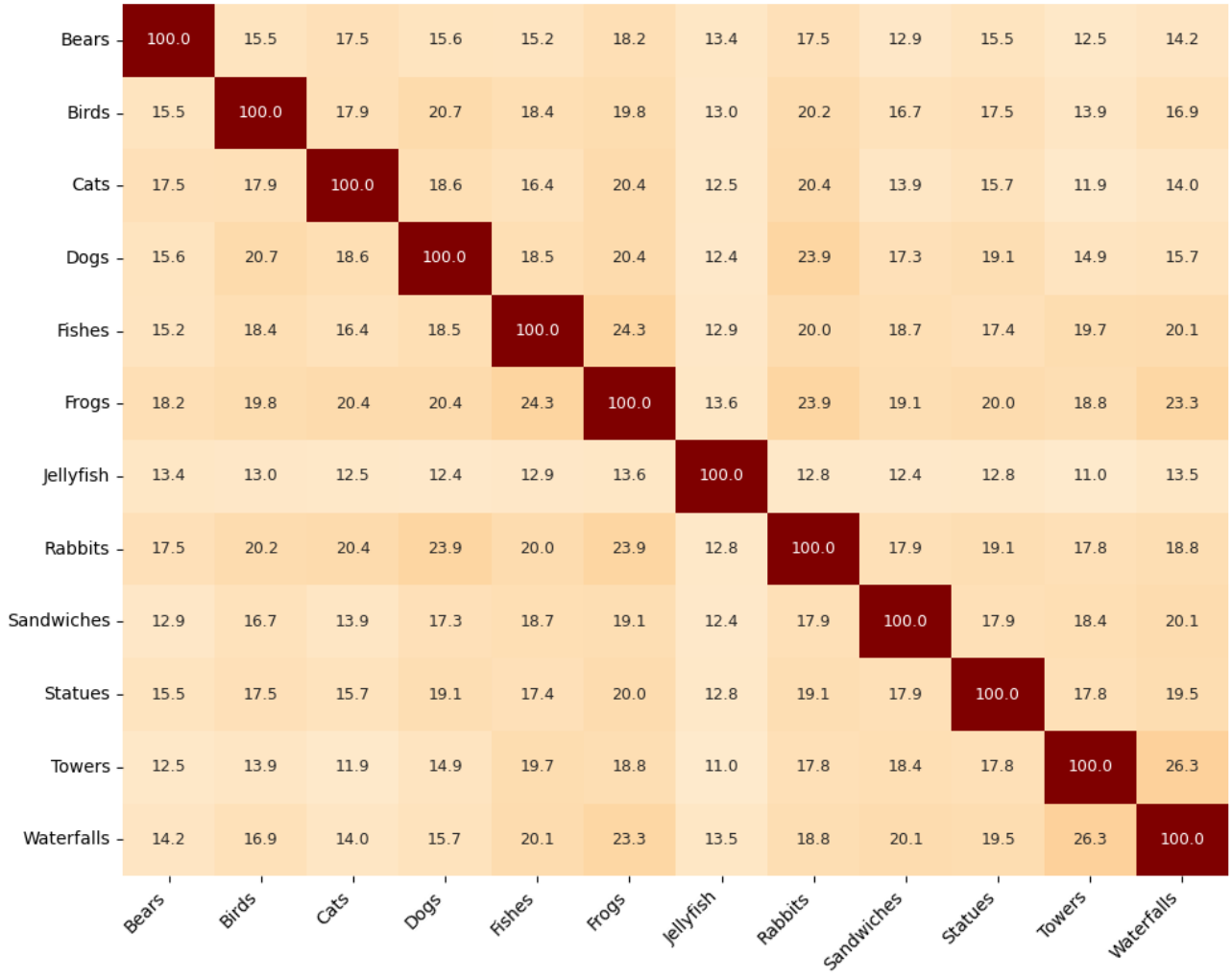

Figure 11. Parameter overlap matrix for CA object selective fine-tuning showing incremental weight changes between consecutive unlearning steps. The reduced overlap demonstrates that selective fine-tuning identifies disjoint sets of important parameters for each concept, minimizing repeated modifications to the same parameters.

C.4. Simultaneous

Simultaneous unlearning experiments preserved the hyperparameter configurations from independent unlearning but consistently reinitialized from the original base model for each unlearning operation. This approach required careful determination of the optimal training duration for effectively unlearning multiple concepts concurrently. We implemented an adaptive stopping criterion where every 250 steps, we performed lightweight evaluation by sampling 15 images per concept and assessing unlearning accuracy. Training continued until the sampled UA exceeded 99%, at which point we stopped and utilized the checkpoint for comprehensive evaluation.

Implementation details varied slightly between methods: ESD maintained its original batch size of 1, with each iteration randomly selecting from the pool of concepts targeted for unlearning. CA implemented multi-unlearning by constructing a combined dataset containing the appropriate anchor pairs for each target concept. Using a batch size of 4, each example randomly selected a target-anchor pair from this combined dataset, enabling parallel unlearning of multiple concepts.

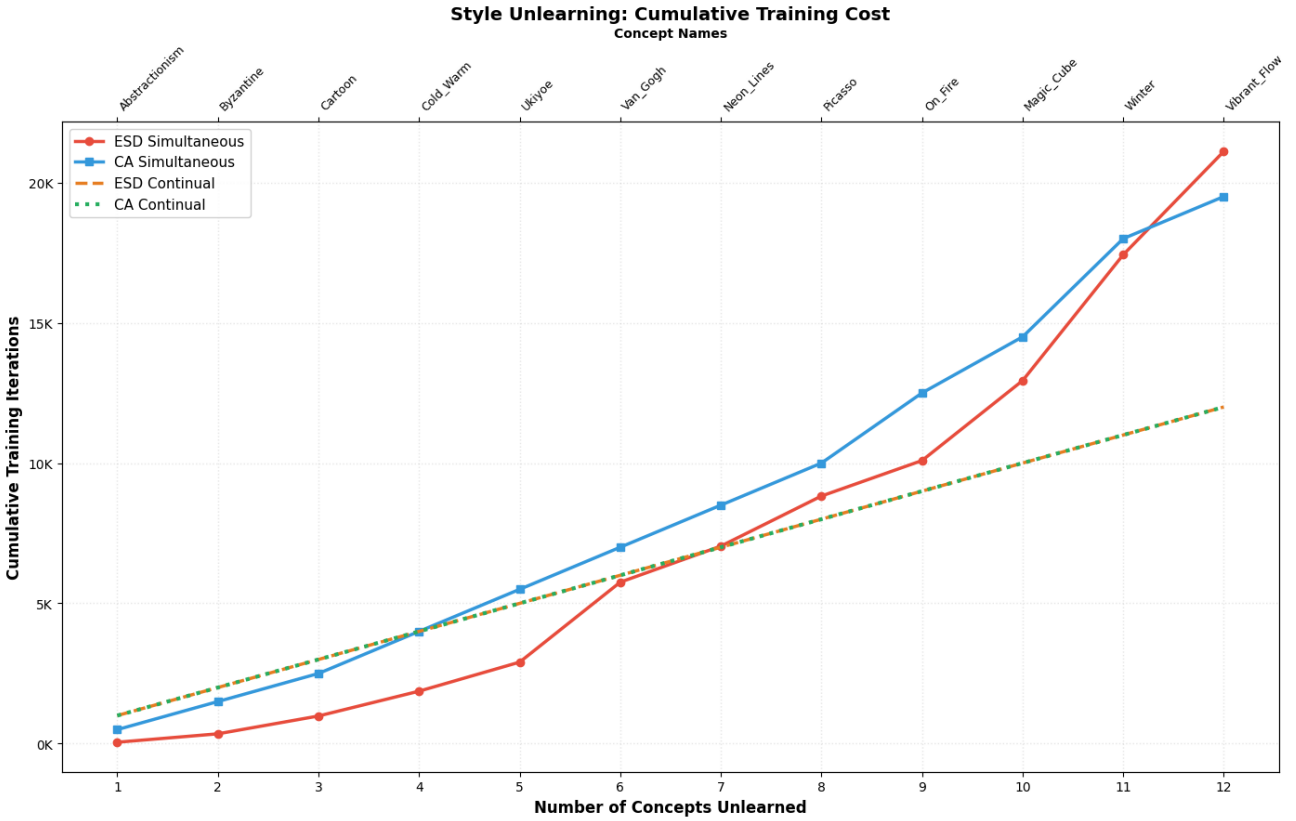

Figure 12. Cumulative training iterations for style concept unlearning comparing simultaneous and continual learning. Continual learning methods (dashed and dotted lines) require a fixed 1000 iterations per concept, resulting in linear growth. Simultaneous learning methods (solid lines) train individual models with variable iteration requirements, leading to non-linear cumulative costs.

C.5. L1

L1 regularization experiments built upon the continual unlearning foundation by incorporating an additional L1 loss term with an empirically tuned scaling hyperparameter. The implementation stored original parameter values before initiating training from the previously unlearned model. During optimization, the sum of absolute parameter differences was calculated, scaled by the loss coefficient, and added to the primary unlearning objective.

Hyperparameter selection involved testing multiple scaling values for the first concept unlearning, with the constraint that UA remained above 90% whenever possible to ensure fair comparison. The final configuration used an L1 weight of 100 for both ESD Style and CA Style. For object unlearning, ESD Class used a weight of 100, while CA Class required a lower weight of 10 to maintain effective unlearning capabilities. ESD Celebrity used a weight of 0.5, and CA Celebrity used a weight of 0.1.

C.6. L2

Similar to the L1 approach, L2 regularization extended continual unlearning by incorporating a squared parameter difference penalty. Implementation involved storing original parameter values before training and computing the sum of squared differences at each step, scaled by a tuned hyperparameter, before adding to the unlearning loss.

Hyperparameter selection followed the same protocol as with L1 regularization, testing multiple values and selecting configurations that maintained UA above 90%. The final implementation used substantially different scaling factors across methods: ESD Style required a weight of 300,000, while CA Style used 75,000. For object unlearning, ESD Class used 75,000 and CA Class used 25,000, reflecting differing sensitivities to parameter perturbation across methodologies. ESD

Celebrity used a weight of 1000 and CA Celebrity used a weight of 50.

C.7. SelFT

Selective finetuning (SelFT) introduced parameter-specific training constraints while maintaining the core continual unlearning framework. The implementation involved a sophisticated parameter importance estimation procedure: for each unlearning method, we computed one forward pass (50 denoising steps) and calculated unlearning losses at each timestep, with gradient tracking enabled only for parameters targeted by the specific algorithm (cross-attention for ESD Style, all except cross-attention for ESD Objects, and KV weights of cross-attention for both CA variants).

For CA, this process involved sampling a noisy latent and calculating predicted noise conditioned under both target and anchor concepts at each timestep. The L2 difference between these predictions was computed with stop gradient applied to the anchor-conditioned prediction. Gradients were stored, and the anchor-conditioned noise applied to the noisy latent before proceeding to the next timestep across 50 DDIM steps. ESD followed a similar procedure but compared target-conditioned predictions against negative classifier-free guidance predictions.

After gradient computation, importance scores were derived by multiplying gradients by parameter weights and taking the absolute value. These scores were flattened, and the top k% of parameter elements selected. A binary mask (1 for selected, 0 for non-selected parameters) was constructed and applied as a gradient hook during training, effectively constraining updates to the most important k% of parameters.

Hyperparameter selection followed our established protocol of testing multiple values and selecting configurations maintaining UA above 90%. The final implementation used a top-k percentage of 10% for both ESD Style and ESD Object, while CA Style and CA Object both used 5%, reflecting differences in parameter update sparsity requirements across methods. ESD and CA Celebrity both used a top-k percentage of 40%.

C.8. TIES Merging

TIES Merging applied a model merging strategy to independently trained unlearning checkpoints. For each unlearning sequence, we conducted a grid search across 4 lambda options [1.25, 1.50, 1.75, 2.0] and 4 top-k values [0.20, 0.40, 0.60, 0.80], yielding 16 candidate models per sequence. Model selection employed a two-stage process: first filtering candidates with UA above 95% to ensure effective unlearning, then selecting among this filtered set the model with the highest average of IRA and CRA metrics, optimizing for both retention capabilities.

D. Detailed Related Work

D.1. From Continual Learning to Continual Unlearning

Continual learning focuses on enabling models to acquire new knowledge incrementally without forgetting previously learned information—a phenomenon known as catastrophic forgetting (Mai et al., 2022; Wang et al., 2024). Existing approaches to mitigate forgetting in continual learning can broadly be classified into four categories: (1) *regularization-based methods*, which incorporate explicit regularization terms to constrain parameter updates (Kirkpatrick et al., 2017; Zenke et al., 2017); (2) *replay-based methods*, which either store a limited set of previous examples in memory buffers (Mai et al., 2021) or employ generative models to synthesize replay samples (Shin et al., 2017); (3) *optimization-based methods*, which directly manipulate optimization procedures through techniques such as gradient projection (Chaudhry et al., 2018) or meta-learning (Javed & White, 2019); and (4) *architecture-based methods*, which introduce task-specific adaptive parameters to the model (Mallya et al., 2018).

Although continual unlearning fundamentally differs from continual learning, key concepts from continual learning remain valuable and adaptable (Heng & Soh, 2023). In this work, we leverage ideas inspired by regularization-based methods from continual learning, introducing L1/L2 regularization baselines. Additionally, while selective parameter updates appear in both paradigms, continual learning methods update the least important parameters to preserve prior knowledge (Mazumder et al., 2021). In contrast, our proposed Selective Fine-Tuning (SelFT) approach identifies and updates the most significant parameters to facilitate effective unlearning.

By bridging insights from continual learning to continual unlearning, our research sets the stage for future investigations. We encourage subsequent studies to further integrate and refine continual learning strategies to address the nuanced challenges

of continual unlearning effectively.

D.2. Selective Fine-tuning

Selecting the most important parameters within a model for a specific task has been extensively investigated for different purposes. To enhance time and memory efficiency, weight pruning methods commonly utilize gradient-based metrics to quantify parameter importance, enabling the removal of redundant parameters (Lee et al., 2018; Molchanov et al., 2016; Tanaka et al., 2020). A similar concept underlies model editing techniques, which aim to precisely locate and alter specific knowledge within a model by directly modifying relevant weights (Dai et al., 2021; Patil et al., 2023; De Cao et al., 2021). Recent work has extended these ideas to unlearning in diffusion models (Fan et al., 2023; Nguyen et al., 2024). Our findings demonstrate that incorporating selective fine-tuning into existing unlearning methodologies significantly enhances their performance in continual unlearning scenarios.

D.3. Model Merging

Early research on model merging focused on averaging parameters of multiple models trained with varied hyperparameters on identical datasets to enhance generalization (Wortsman et al., 2022). Concurrently, this strategy has been extended to multi-task learning, where models trained on diverse vision tasks have their weights averaged to achieve improved performance (Matena & Raffel, 2022; Ilharco et al., 2022). Since then, numerous advanced methods have emerged to refine the basic merging approach (fine-tuning followed by merging), including linearized fine-tuning (Ortiz-Jimenez et al., 2023), sparsifying task vectors(Davari & Belilovsky, 2024; Yu et al., 2024), and selectively merging subsets of weights(Yadav et al., 2023; Stoica et al., 2023).

Recent concurrent studies have also explored model merging techniques specifically tailored for unlearning in large language models (LLMs) (Kuo et al., 2025; Kadhe et al., 2024). However, to the best of our knowledge, this paper presents the **first** exploration of model merging for unlearning within the context of text-to-image generation.

E. Detailed Preliminary

E.1. Text-to-image generative models

Diffusion-based generative models create data by gradually removing noise from an initial Gaussian sample. Sampling begins with $x_T \sim \mathcal{N}(0, \mathbf{I})$ and proceeds for T discrete steps; at each step t, a neural network estimates the noise component ϵ_t contained in the current state x_t and subtracts it to obtain the next, cleaner state x_{t-1}. The reverse trajectory $\{x_T, \ldots, x_0\}$ therefore forms a first-order Markov chain whose probability factorises as

$$p_\theta(x_{T:0}) \;=\; p(x_T) \prod_{t=T}^{1} p_\theta\big(x_{t-1} \,\big|\, x_t\big), \tag{1}$$

with x_0 representing the final, fully denoised image.

Latent Diffusion Models (LDMs) apply the same denoising principle in a lower-dimensional latent space to reduce memory and compute demands. A fixed variational autoencoder encodes an image x as $z = \mathcal{E}(x)$; Gaussian noise is then incrementally added to this latent to produce a sequence $\{z_t\}$. A single network, parameterized by θ, learns to predict the injected noise $\epsilon_\theta(z_t, c, t)$ at each timestep t, optionally conditioned on an auxiliary text embedding c. Training minimizes the mean-squared error

$$\mathcal{L} \;=\; \mathbb{E}_{z_t \sim \mathcal{E}(x),\, t,\, c,\, \epsilon \sim \mathcal{N}(0,1)} \Big[\big\| \epsilon - \epsilon_\theta(z_t, c, t) \big\|_2^2 \Big]. \tag{2}$$

After denoising back to a clean latent z_0, the decoder $\mathcal{D}$ reconstructs the final image via $x_0 \approx \mathcal{D}(z_0)$.

E.2. Machine Unlearning for Generative Models

Modern pre-trained generative models, denoted by G_θ, typically support text-to-image generation. Given a textual prompt a, the model first encodes it into $c(a)$, which is then used to guide the diffusion process during image synthesis: $I = G_\theta\left(c(a)\right)$. Ideally, if the textual prompt contains a concept q, *i.e.*, $q \in a$, the generated image I should accurately reflect it. One approach to assess it is to input I into a recognition model f_ϕ, such as CLIP (Radford et al., 2021), and check whether $f_\phi(I)$ aligns with q.

The goal of unlearning for generative models is to remove the model's ability to generate certain targeted concepts in its output images. Let $q^\star$ denote such a concept, and let $\theta(q^\star)$ denote the model after unlearning $q^\star$. For a generated image $I = G_{\theta(q^\star)}(c(a))$ where the prompt contains $q^\star$, i.e., $q^\star \in a$, we should have $f_\phi(I) \neq q^\star$. For all other concepts $q \neq q^\star$, the model should retain them; that is, if $q \in a$, then we should have $f_\phi(I) = q$.

F. Detailed Setup

F.1. Detailed Data Setup

Previous works on unlearning concepts in diffusion models lacked standardized evaluation protocols, using inconsistent metrics ranging from CLIP Score similarity (Wu et al., 2025; Kumari et al., 2023; Gandikota et al., 2023) to subjective human evaluations (Gandikota et al., 2023), making method comparison difficult. To address this, we adopt UNLEARN-CANVAS (Zhang et al., 2024b) as our evaluation framework. It provides a fine-tuned Stable Diffusion checkpoint and specialized classifiers trained on 60 artistic styles and 20 object categories. The checkpoint guarantees all selected concepts can be strongly generated initially (>98% top-1 accuracy), while the classifiers offer robust, objective evaluation metrics. Our methodology works as follows: when unlearning a concept (e.g., "Van Gogh" style), we generate diverse images in that style post-unlearning and evaluate the classifier's detection ability. The classifier's error rate accurately represents unlearning success. This framework measures three critical aspects: (1) unlearning accuracy - reduced classifier detection of the target concept; (2) in-domain retention accuracy - continued accuracy on other concepts of the same type; and (3) cross-domain retention accuracy - performance on concepts from different domains. This provides a standardized foundation for comparing unlearning methods.

F.2. Detailed Unlearning Sequence Setup

To systematically evaluate continual unlearning performance, we establish two distinct experimental settings that mirror realistic usage patterns.

Sequential Style Unlearning: In this setting, we simulate a scenario where artistic styles are progressively removed from the model's generation capabilities. We construct a random sequence of 12 unique artistic styles to be unlearned sequentially, with each style representing a distinct unlearning request received over time. To comprehensively assess the model's retention capabilities, we maintain a held-out evaluation set comprising 12 additional styles and 8 objects that remain untargeted throughout the unlearning sequence. This held-out set enables us to measure both in-domain retention (the model's ability to generate other artistic styles) and cross-domain retention (the model's continued proficiency in object generation).

Sequential Object Unlearning: Complementing the style-focused setting, we evaluate continual unlearning in the object domain, where specific object categories are sequentially removed. Similar to the style setting, we randomly select 12 unique objects to form the unlearning sequence. Importantly, we utilize the identical held-out evaluation set employed in style unlearning, ensuring fair comparison across domains and eliminating potential biases introduced by different evaluation sets.

By tracking performance across both unlearned concepts and held-out sets throughout the unlearning sequence, we can quantify how each sequential unlearning step affects the model's overall capabilities.

F.3. Detailed Unlearning Methods Setup

Below, we describe two pioneering unlearning methods evaluated in our benchmark: **Erasing Concepts from Diffusion Models (ESD)** (Gandikota et al., 2023) and **Concept Ablation (CA)** (Kumari et al., 2023). These methods represent distinct paradigms for unlearning in diffusion models.

ESD leverages classifier-free guidance mechanics to steer generation away from a target concept by reversing the guidance direction. The method exploits the model's own knowledge of concepts, eliminating the need for external datasets. For lightweight removal tasks—such as erasing artistic styles—ESD updates only the cross-attention layers (ESD-x), which serve as gateways for text conditioning in the image generation process and activate when certain tokens are present in the text prompt. For concepts more deeply embedded throughout the model, such as object erasure, ESD updates all non-cross-attention parameters (ESD-u) to achieve global erasure independent of text conditioning. The training objective minimizes the difference between the model's prediction and a negatively guided target:

$$\epsilon_\theta(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) - \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)]$$

where θ^* represents the frozen original model and η controls the guidance strength. This formulation encourages the model to reduce the likelihood of generating images aligned with the target concept c.

CA approaches unlearning through concept substitution, learning to match the image distribution of a target concept to an anchor concept (e.g., "Van Gogh painting" $\rightarrow$ "painting"). Unlike ESD's broad parameter updates, CA specifically targets the interaction between text embeddings and latent image features by updating only the key and value projection matrices in the UNet's cross-attention layers. This targeted approach modifies how text conditions are integrated with visual features during the denoising process, effectively redirecting the model's attention from target concept tokens to broader anchor concept representations. The training objective minimizes the discrepancy between noise predictions for target and anchor concepts:

$$\mathcal{L}_{\text{model}}(\mathbf{x}, \mathbf{c}, \mathbf{c}^*) = \mathbb{E}_{\epsilon,\mathbf{x},\mathbf{c}^*,\mathbf{c},t}\left[w_t \left\|\hat{\Phi}(x_t, \mathbf{c}, t).\text{sg}() - \hat{\Phi}(x_t, \mathbf{c}^*, t)\right\|\right]$$

where $\mathbf{c}^*$ denotes the anchor concept and $\text{sg}()$ applies stop-gradient to prevent the anchor concept from being altered. CA also incorporates a regularization term using pre-generated anchors to maintain the anchor concept's original characteristics while redirecting target concept generation.

Together, these methods offer complementary approaches to unlearning—ESD focusing on guidance-based distribution reshaping through reversed classifier-free guidance, and CA emphasizing controlled modification of text-image cross-attention interactions to achieve semantic concept substitution.

G. Qualitative Results

To illustrate the qualitative meaning of IRA and CRA metrics, we present example style-object pairs from the held-out sets used in their computation. Ideally, these generated images should remain visually consistent with those from the base model, regardless of the number of unlearning requests. In Figure 13, we compare continual unlearning with our proposed baselines for CA style unlearning. Figure 14 further shows results for ESD object unlearning.

Table 1. ESD Independent Style and CA Independent Style

Style	ESD Independent Style			CA Independent Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	100.00	73.96	93.27	100.00	67.08	93.65
Byzantine	100.00	94.58	90.96	100.00	88.33	90.38
Cartoon	100.00	78.75	94.23	100.00	65.63	92.12
Cold_Warm	100.00	95.63	92.88	100.00	87.08	96.35
Ukiyoe	100.00	85.63	92.31	100.00	97.08	94.62
Van_Gogh	100.00	73.54	96.73	100.00	81.88	88.85
Neon_Lines	100.00	81.46	95.19	100.00	71.46	94.42
Picasso	100.00	84.79	95.96	100.00	79.58	94.23
On_Fire	100.00	96.46	90.19	100.00	80.42	91.35
Magic_Cube	100.00	86.88	94.23	100.00	81.04	90.77
Winter	100.00	97.92	93.65	100.00	90.21	89.62
Vibrant_Flow	100.00	71.25	95.38	100.00	37.08	90.77

Table 2. ESD Continual Style and CA Continual Style

Style	ESD Continual Style			CA Continual Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	100.00	73.96	93.27	100.00	67.08	93.65
Byzantine	100.00	40.21	79.82	100.00	37.08	91.25
Cartoon	97.50	22.08	67.83	100.00	14.38	68.17
Cold_Warm	100.00	8.33	57.81	100.00	24.79	53.91
Ukiyoe	99.00	3.96	40.15	100.00	34.79	49.56
Van_Gogh	98.33	3.33	26.11	100.00	37.92	49.44
Neon_Lines	100.00	3.12	17.11	100.00	16.46	36.18
Picasso	99.06	2.71	10.50	100.00	22.71	40.13
On_Fire	95.83	3.75	10.60	100.00	15.83	31.79
Magic_Cube	91.00	2.71	9.66	100.00	11.25	29.55
Winter	94.77	0.63	9.13	100.00	9.38	34.24
Vibrant_Flow	98.12	0.00	9.48	100.00	5.21	32.40

Table 3. Comparison of ESD Simultaneous Style and CA Simultaneous Style Metrics

Style	ESD Simultaneous Style			CA Simultaneous Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	97.50	75.90	95.45	100.00	73.12	94.23
Byzantine	95.00	70.63	93.81	100.00	63.96	94.29
Cartoon	93.61	70.28	96.61	98.33	45.21	96.50
Cold_Warm	98.34	77.09	92.97	98.75	32.50	86.72
Ukiyoe	99.17	68.75	92.99	99.50	35.00	95.59
Van_Gogh	99.58	66.04	88.80	100.00	35.62	95.97
Neon_Lines	97.50	56.81	90.66	99.64	30.21	96.84
Picasso	96.25	57.16	85.29	99.38	26.46	90.00
On_Fire	98.24	45.28	89.40	99.72	23.75	85.24
Magic_Cube	98.17	34.79	57.69	99.50	21.46	88.41
Winter	98.71	50.91	79.20	98.64	17.71	55.33
Vibrant_Flow	98.96	45.83	86.84	99.17	9.17	55.21

Table 4. Comparison of ESD Continual L1 Style and CA Continual L1 Style Metrics

Style	ESD Continual L1 Style			CA Continual L1 Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	95.00	70.83	96.73	100.00	93.13	95.19
Byzantine	96.25	34.79	91.07	90.00	79.58	96.61
Cartoon	98.33	22.71	83.33	92.50	53.54	96.83
Cold_Warm	96.88	20.83	65.31	94.38	48.12	97.03
Ukiyoe	94.00	13.75	41.47	98.00	40.00	96.62
Van_Gogh	97.92	7.29	31.67	95.42	39.79	97.50
Neon_Lines	100.00	5.00	28.03	99.64	25.00	97.24
Picasso	99.06	4.38	26.50	98.44	28.33	95.25
On_Fire	99.44	5.21	25.71	99.17	23.75	96.19
Magic_Cube	99.50	3.96	24.09	99.50	20.42	96.59
Winter	99.32	4.58	25.00	98.41	24.58	94.67
Vibrant_Flow	99.58	4.38	22.29	98.54	13.54	93.65

Table 5. Comparison of ESD Continual L2 Style and CA Continual L2 Style Metrics

Style	ESD Continual L2 Style			CA Continual L2 Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	92.50	75.00	96.54	97.50	88.96	95.19
Byzantine	92.50	44.58	88.39	87.50	75.42	94.46
Cartoon	96.67	30.42	81.33	97.50	43.54	95.33
Cold_Warm	98.12	26.04	71.72	97.50	40.62	90.16
Ukiyoe	93.50	21.25	59.85	97.00	37.08	94.41
Van_Gogh	95.42	9.79	50.69	93.33	36.04	90.42
Neon_Lines	96.79	6.25	42.89	98.21	21.25	89.74
Picasso	99.38	6.67	38.00	97.81	23.75	88.25
On_Fire	99.17	6.67	33.33	97.50	21.25	84.05
Magic_Cube	99.75	5.00	30.57	98.75	13.54	79.43
Winter	98.86	6.88	27.61	97.95	19.37	68.80
Vibrant_Flow	99.58	4.79	26.15	99.79	12.29	48.33

Table 6. Comparison of ESD Continual SelFT Style and CA Continual SelFT Style Metrics

Style	ESD Continual SelFT Style			CA Continual SelFT Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	100.00	60.62	97.31	100.00	95.21	95.38
Byzantine	98.75	28.12	91.79	97.50	89.17	94.82
Cartoon	98.33	20.00	86.50	95.83	71.46	96.00
Cold_Warm	98.12	17.92	76.88	98.12	67.92	96.41
Ukiyoe	95.50	11.67	65.74	98.50	58.12	97.06
Van_Gogh	99.58	5.83	54.72	99.17	51.04	96.94
Neon_Lines	99.64	6.25	46.71	100.00	29.17	96.18
Picasso	99.69	6.67	41.25	99.69	34.17	95.00
On_Fire	99.44	5.42	33.21	99.72	28.33	90.12
Magic_Cube	99.25	3.54	28.64	100.00	18.54	90.11
Winter	99.77	4.38	25.33	98.18	24.38	90.76
Vibrant_Flow	99.79	2.08	20.31	99.79	12.71	89.58

Table 7. Comparison of ESD TIES Merge Style and CA TIES Merge Style Metrics

Style	ESD TIES Merge Style			CA TIES Merge Style		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Abstractionism	100.00	73.96	93.27	100.00	67.08	93.65
Byzantine	96.25	81.04	90.54	100.00	55.21	88.75
Cartoon	96.67	57.92	87.67	97.50	45.63	90.33
Cold_Warm	96.88	52.50	87.66	96.88	45.21	89.53
Ukiyoe	96.00	47.92	84.41	96.50	37.50	91.62
Van_Gogh	97.50	39.79	82.78	97.08	29.38	89.31
Neon_Lines	96.43	46.04	86.97	98.21	27.50	88.95
Picasso	96.88	41.87	84.88	98.75	22.71	86.75
On_Fire	95.56	41.04	83.57	96.39	20.00	87.62
Magic_Cube	96.25	35.83	83.30	95.75	23.13	83.64
Winter	95.23	36.67	80.87	95.23	22.50	83.91
Vibrant_Flow	98.75	24.38	96.87	97.08	15.00	87.08

Table 8. Comparison of ESD Independent Object and CA Independent Object Metrics

Object	ESD Independent Object			CA Independent Object		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	100.00	68.75	86.48	96.67	77.50	99.63
Birds	100.00	61.04	87.59	100.00	91.67	98.70
Cats	100.00	40.62	79.63	100.00	82.50	99.81
Dogs	95.00	74.79	99.07	100.00	70.63	99.81
Fishes	96.67	82.92	86.11	100.00	87.50	96.48
Frogs	96.67	67.29	87.04	100.00	84.38	98.70
Jellyfish	95.00	82.29	88.52	100.00	84.79	98.70
Rabbits	96.67	62.29	89.26	100.00	83.54	99.07
Sandwiches	100.00	73.33	60.00	100.00	85.63	98.89
Statues	98.33	80.83	89.63	100.00	86.88	99.81
Towers	95.00	85.00	87.96	100.00	91.67	99.26
Waterfalls	86.67	92.08	91.48	100.00	76.25	98.89

Table 9. Comparison of ESD Object Continual and CA Object Continual Metrics

Object	ESD Object Continual			CA Object Continual		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	100.00	68.75	86.48	96.67	77.50	99.63
Birds	100.00	40.83	66.33	100.00	77.71	98.67
Cats	100.00	24.38	35.30	98.33	61.67	99.39
Dogs	100.00	19.37	26.67	99.58	43.96	97.50
Fishes	100.00	14.79	17.69	97.33	36.67	95.90
Frogs	100.00	12.29	9.17	99.44	33.33	92.14
Jellyfish	100.00	12.50	8.11	96.90	36.04	90.22
Rabbits	100.00	12.50	7.81	98.75	33.75	93.75
Sandwiches	100.00	12.71	7.55	95.93	35.00	90.29
Statues	100.00	12.71	5.00	96.67	40.42	89.26
Towers	100.00	12.71	3.25	97.58	38.75	50.96
Waterfalls	100.00	12.71	2.33	95.83	33.75	37.75

Table 10. Comparison of ESD Simultaneous Object and CA Simultaneous Object Metrics

Object	ESD Simultaneous Object			CA Simultaneous Object		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	100.00	68.75	88.70	93.33	88.33	99.63
Birds	96.67	73.33	93.83	94.72	84.24	99.06
Cats	98.89	54.17	82.88	95.56	78.13	99.09
Dogs_Warm	97.08	52.92	86.25	95.00	71.04	99.35
Fishes	97.00	58.33	88.59	97.67	70.00	98.33
Frogs	98.89	44.79	79.29	98.61	68.05	99.21
Jellyfish	97.86	31.04	61.67	95.64	68.68	99.22
Rabbits	98.96	38.96	64.79	96.04	60.49	98.37
Sandwiches	96.67	33.54	40.20	97.04	56.88	98.01
Statues	99.50	23.75	38.70	97.44	52.99	96.82
Towers	99.17	18.54	33.43	96.51	60.35	97.63
Waterfalls	97.92	27.71	26.42	97.92	56.60	96.75

Table 11. Comparison of ESD Continual L1 Object and CA Continual L1 Object Metrics

Object	ESD Continual L1 Object			CA Continual L1 Object		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	98.33	95.21	94.44	100.00	86.67	100.00
Birds	88.33	82.29	79.83	100.00	86.25	99.00
Cats	93.89	80.00	77.12	98.89	75.42	98.48
Dogs	94.17	70.21	75.00	97.50	58.54	99.31
Fishes	94.00	63.75	70.26	95.00	58.75	98.59
Frogs	99.44	41.88	65.83	97.22	58.12	98.69
Jellyfish	94.05	32.08	53.44	94.29	57.08	98.44
Rabbits	96.67	28.12	46.77	94.79	54.17	99.17
Sandwiches	94.26	26.46	39.90	84.81	60.21	98.73
Statues	97.00	22.71	31.85	82.50	61.67	97.69
Towers	97.27	23.12	24.12	87.42	63.13	97.72
Waterfalls	96.39	22.50	16.58	83.33	61.87	96.83

Table 12. Comparison of ESD Continual L2 Object and CA Continual L2 Object Metrics

Object	ESD Continual L2 Object			CA Continual L2 Object		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	98.33	92.92	91.48	100.00	89.17	99.81
Birds	80.83	79.17	81.50	96.67	88.54	99.00
Cats	92.78	68.96	72.42	99.44	83.33	98.79
Dogs	92.50	61.25	60.69	97.92	70.83	99.44
Fishes	91.00	42.92	50.00	88.33	66.25	98.46
Frogs	92.78	41.67	44.05	96.39	62.92	98.93
Jellyfish	93.10	34.79	37.67	92.86	65.00	99.00
Rabbits	93.75	32.92	36.56	88.12	57.29	99.79
Sandwiches	90.56	33.33	32.06	75.00	72.50	99.61
Statues	94.17	26.25	27.04	63.83	73.96	99.35
Towers	94.55	21.46	22.54	62.73	71.67	99.21
Waterfalls	94.17	21.25	17.67	59.44	67.08	99.08

Table 13. Comparison of ESD Continual SelFT Object and CA Continual SelFT Object Metrics

Object	ESD Continual SelFT Object			CA Continual SelFT Object		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	91.67	93.54	99.63	100.00	90.83	99.63
Birds	94.17	74.79	93.83	96.67	92.08	99.33
Cats	96.67	62.92	89.24	99.44	82.92	99.55
Dogs	97.08	59.58	82.78	99.17	71.04	99.58
Fishes	94.00	41.25	72.05	98.67	70.21	96.79
Frogs	97.78	31.88	64.64	100.00	69.38	95.24
Jellyfish	98.33	21.25	43.89	99.76	60.42	90.89
Rabbits	99.38	17.29	29.90	99.79	58.96	96.04
Sandwiches	99.26	18.33	13.53	98.52	59.79	95.10
Statues	100.00	16.04	11.48	96.00	69.17	95.46
Towers	99.55	13.75	5.88	97.12	68.54	95.96
Waterfalls	99.86	13.54	3.25	97.64	65.62	94.67

Table 14. Comparison of ESD TIES Merge Object and CA TIES Merge Object Metrics

Object	ESD TIES Merge Object			CA TIES Merge Object		
	UA (%)	IRA (%)	CRA (%)	UA (%)	IRA (%)	CRA (%)
Bears	100.00	68.75	86.48	96.67	77.50	99.63
Birds	95.83	55.83	75.83	99.17	76.67	98.50
Cats	96.11	42.50	67.73	98.33	66.88	98.50
Dogs	95.42	55.00	74.31	97.92	64.38	99.44
Fishes	96.00	44.38	52.56	96.00	54.17	98.59
Frogs	95.83	43.96	49.40	97.22	58.75	98.57
Jellyfish	95.71	26.88	34.00	95.95	37.50	97.44
Rabbits	95.00	28.96	34.27	95.83	42.50	98.75
Sandwiches	95.37	28.54	27.35	96.48	28.33	95.59
Statues	95.83	25.62	22.59	96.50	31.04	95.37
Towers	95.76	23.75	18.25	95.61	34.17	95.26
Waterfalls	95.83	23.54	17.42	96.67	24.58	92.83

Table 15. Comparison of ESD Independent Celebrity and CA Independent Celebrity Metrics

Celebrity	ESD Independent Celebrity		CA Independent Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	100.00	$88.21_{(92.13)}$	98.24	$92.41_{(96.51)}$
Benicio_Del_Toro	100.00	$89.18_{(93.14)}$	100.00	$92.23_{(96.32)}$
Aziz_Ansari	99.57	$91.14_{(95.19)}$	100.00	$93.61_{(97.76)}$
Oprah_Winfrey	99.57	$81.52_{(85.14)}$	100.00	$81.85_{(85.48)}$
Betty_White	100.00	$82.57_{(86.23)}$	99.09	$85.80_{(89.60)}$
Megan_Fox	100.00	$92.23_{(96.32)}$	97.07	$92.22_{(96.31)}$

Table 16. Comparison of ESD Continual Celebrity and CA Continual Celebrity Metrics

Celebrity	ESD Continual Celebrity		CA Continual Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	100.00	$88.21_{(92.13)}$	98.24	$92.41_{(96.51)}$
Benicio_Del_Toro	99.56	$71.29_{(74.45)}$	99.50	$87.35_{(91.23)}$
Aziz_Ansari	95.87	$62.56_{(65.34)}$	99.58	$83.14_{(86.83)}$
Oprah_Winfrey	95.49	$41.59_{(43.43)}$	99.83	$58.17_{(60.75)}$
Betty_White	92.66	$25.16_{(26.28)}$	98.92	$32.73_{(34.18)}$
Megan_Fox	94.86	$14.93_{(15.59)}$	98.74	$19.33_{(20.19)}$

Table 17. Comparison of ESD Simultaneous Celebrity and CA Simultaneous Celebrity Metrics

Celebrity	ESD Simultaneous Celebrity		CA Simultaneous Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	99.54	$89.39_{(93.36)}$	100.00	$91.31_{(95.36)}$
Benicio_Del_Toro	99.13	$80.48_{(84.05)}$	97.78	$89.52_{(93.49)}$
Aziz_Ansari	99.13	$81.84_{(84.47)}$	99.19	$87.38_{(91.26)}$
Oprah_Winfrey	99.05	$69.61_{(72.70)}$	99.34	$71.25_{(74.41)}$
Betty_White	99.21	$59.01_{(61.63)}$	97.63	$56.95_{(59.48)}$
Megan_Fox	99.13	$52.64_{(54.98)}$	98.76	$51.98_{(54.29)}$

Table 18. Comparison of ESD Continual L1 Celebrity and CA Continual L1 Celebrity Metrics

Celebrity	ESD Continual L1 Celebrity		CA Continual L1 Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	99.53	$86.11_{(89.93)}$	85.88	$91.46_{(95.52)}$
Benicio_Del_Toro	84.93	$71.38_{(74.55)}$	88.50	$89.72_{(93.70)}$
Aziz_Ansari	84.52	$63.09_{(65.89)}$	93.77	$88.09_{(92.00)}$
Oprah_Winfrey	85.44	$44.90_{(46.89)}$	96.54	$72.45_{(75.67)}$
Betty_White	83.53	$28.07_{(29.32)}$	96.55	$50.46_{(52.70)}$
Megan_Fox	77.31	$25.70_{(26.84)}$	94.61	$37.25_{(38.90)}$

Table 19. Comparison of ESD Continual L2 Celebrity and CA Continual L2 Celebrity Metrics

Celebrity	ESD Continual L2 Celebrity		CA Continual L2 Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	98.06	$86.72_{(90.57)}$	95.21	$92.55_{(96.66)}$
Benicio_Del_Toro	90.31	$72.95_{(76.19)}$	99.51	$88.22_{(92.14)}$
Aziz_Ansari	87.61	$63.60_{(66.42)}$	99.61	$84.98_{(88.75)}$
Oprah_Winfrey	86.41	$43.86_{(45.81)}$	99.65	$64.40_{(67.29)}$
Betty_White	85.92	$29.69_{(31.01)}$	98.84	$37.75_{(39.43)}$
Megan_Fox	86.68	$4.59_{(4.79)}$	97.34	$22.96_{(23.98)}$

Table 20. Comparison of ESD Continual SelFT Celebrity and CA Continual SelFT Celebrity Metrics

Celebrity	ESD Continual SelFT Celebrity		CA Continual SelFT Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	97.78	$85.76_{(89.57)}$	97.40	$91.35_{(95.40)}$
Benicio_Del_Toro	89.56	$65.49_{(68.40)}$	95.66	$86.39_{(90.22)}$
Aziz_Ansari	92.53	$54.30_{(56.71)}$	97.50	$84.01_{(87.74)}$
Oprah_Winfrey	86.94	$36.58_{(38.20)}$	99.24	$61.98_{(64.73)}$
Betty_White	86.92	$20.04_{(20.93)}$	98.17	$37.45_{(39.11)}$
Megan_Fox	82.23	$6.16_{(6.43)}$	96.25	$26.68_{(27.86)}$

Table 21. Comparison of ESD TIES Merge Celebrity and CA TIES Merge Celebrity Metrics

Celebrity	ESD TIES Merge Celebrity		CA TIES Merge Celebrity	
	UA (%)	RA (%)	UA (%)	RA (%)
Neil_Degrasse_Tyson	100.00	$88.21_{(92.13)}$	98.24	$92.41_{(96.51)}$
Benicio_Del_Toro	98.74	$84.99_{(88.76)}$	99.40	$87.84_{(91.74)}$
Aziz_Ansari	98.14	$80.24_{(83.80)}$	98.50	$86.02_{(89.84)}$
Oprah_Winfrey	97.37	$65.22_{(68.11)}$	97.24	$73.08_{(76.32)}$
Betty_White	96.25	$40.95_{(42.77)}$	96.13	$48.63_{(50.79)}$
Megan_Fox	95.92	$34.18_{(35.70)}$	96.02	$33.83_{(35.33)}$

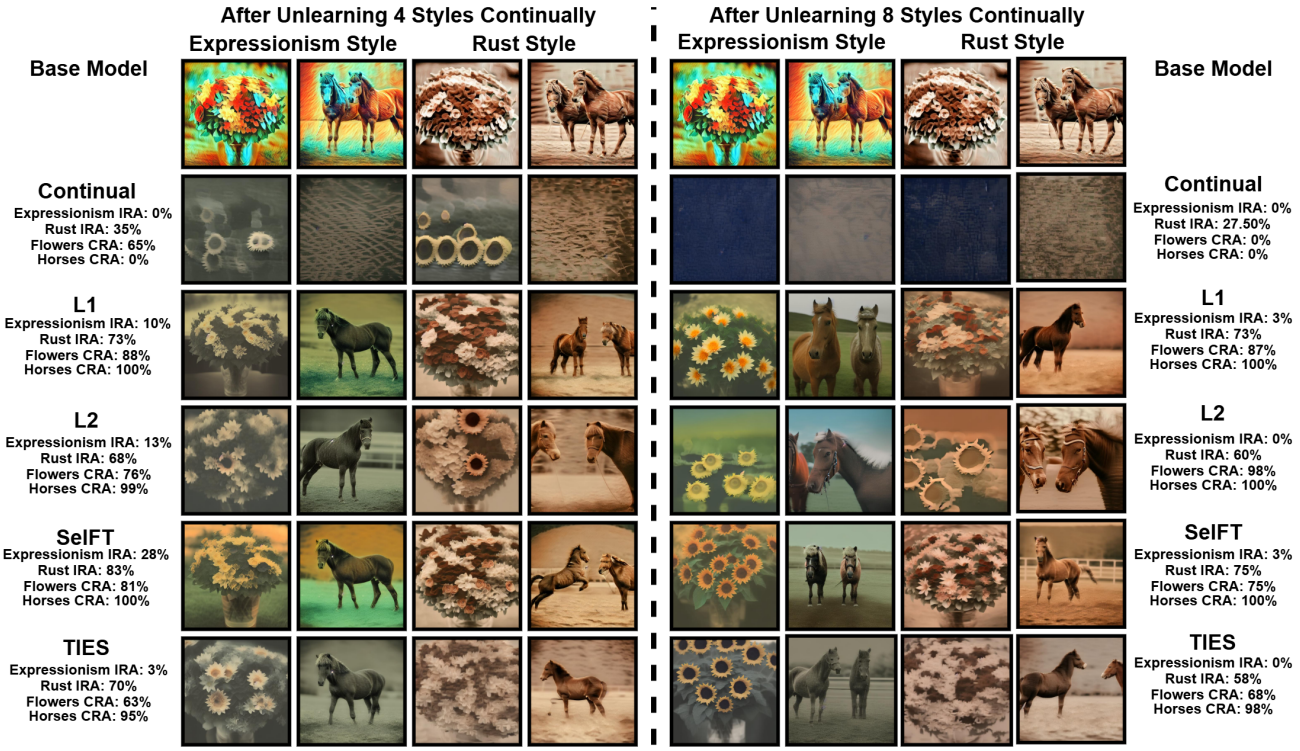

Figure 13. Qualitative comparison of continual unlearning methods using CA after sequentially unlearning 4 styles (left) and 8 styles (right). Generated images show held-out styles (Expressionism, Rust) and held-out objects (flowers, horses) to demonstrate how different baselines (Continual, L1, L2, SelFT, TIES) preserve original model performance compared to continual. Retention accuracies for displayed styles and objects are shown beneath each method.

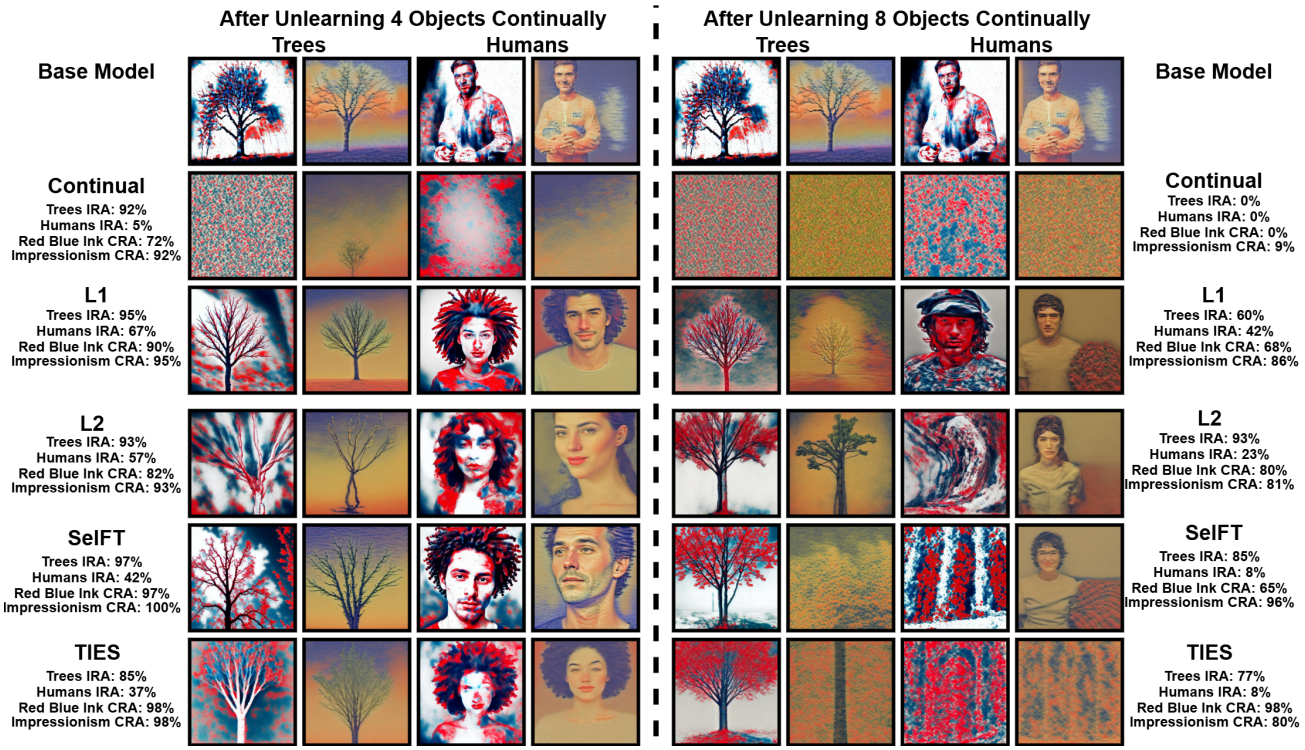

Figure 14. Qualitative comparison of continual unlearning methods using ESD after sequentially unlearning 4 objects (left) and 8 objects (right). Generated images show held-out objects (trees, humans) and held-out styles (Red Blue Ink, Impressionism) to demonstrate how different baselines (Continual, L1, L2, SelFT, TIES) preserve original model performance compared to continual. Retention accuracies for displayed objects and styles are shown beneath each method.