Train Once, Forget Precisely: Anchored Optimization for Efficient Post-Hoc Unlearning

Anonymous Authors[1]

Abstract

As machine learning systems increasingly rely on data subject to privacy regulation, selectively unlearning specific information from trained models has become essential. In image classification, this involves removing the influence of particular training samples, semantic classes, or visual styles without full retraining. We introduce **Forget-Aligned Model Reconstruction (FAMR)**, a theoretically grounded and computationally efficient framework for post-hoc unlearning in deep image classifiers. FAMR frames forgetting as a constrained optimization problem that minimizes a uniform-prediction loss on the forget set while anchoring model parameters to their original values via an ℓ_2 penalty. A theoretical analysis links FAMR's solution to influence-function-based retraining approximations, with bounds on parameter and output deviation. Empirical results on class forgetting tasks using CIFAR-10 and ImageNet-100 demonstrate FAMR's effectiveness, with strong performance retention and minimal computational overhead. The framework generalizes naturally to concept and style erasure, offering a scalable and certifiable route to efficient post-hoc forgetting in vision models.

1. Introduction

As machine learning systems become increasingly pervasive in sensitive domains, such as medical diagnostics and user-facing recommendation engines, ensuring compliance with privacy regulations is paramount. The "right to be forgotten," codified in regulations such as the EU's General Data Protection Regulation (GDPR), mandates that individuals can request deletion of their data and any downstream influence it may have on deployed models. This has led to the emerging research field of *machine unlearning*, which aims to remove specific information—e.g., training samples, semantic concepts, or stylistic patterns—from trained models without requiring full retraining (Cao & Yang, 2015).

In image classification, forgetting a data point, class, or visual concept is particularly challenging due to the distributed and entangled nature of learned representations. Retraining from scratch on the remaining data, while effective, is computationally expensive and often infeasible at scale. As a remedy, various unlearning strategies have been developed to approximate the retrained model's behavior without the associated cost (Zhang et al., 2024b).

Frequent retraining incurs prohibitive latency, particularly for large-scale image classifiers (Gu et al., 2024). Even differentially private training (ε-DP) only bounds contributions in expectation and cannot guarantee complete erasure (Domingo-Ferrer et al., 2021). Early work by Bourtoule *et al.* (2021) introduced the SISA (Sharded, Isolated, Sliced, and Aggregated) framework, which partitions data and retains multiple shard-specific models to facilitate point-wise retraining. Influence-function-based approaches, such as those proposed by Guo *et al.* (2019) and Sekhari *et al.* (2021), estimate the effect of removing specific training samples using a one-step Newton update, although such techniques rely on convex loss assumptions and are often unreliable in deep networks.

Several architecture-based approaches tackle the problem differently. Forsaken (Ma et al., 2023) learns a mask over neurons to erase the influence of forgotten data. In generative modeling, diffusion and transformer-based methods now support object- or identity-style forgetting through fine-tuning or prompt editing (Zhang et al., 2024a). Panda *et al.* (2024) introduced a label-annealing strategy to iteratively erase high-level concepts. However, many of these approaches lack formal guarantees and are typically confined to specific architectures or datatypes. Overall, while machine unlearning is gaining traction, achieving efficient, generalizable, and certifiable forgetting remains a significant challenge.

In this study, we introduce **Forget-Aligned Model Reconstruction (FAMR)**, a post-hoc forgetting framework that

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

directly modifies a trained image classifier to erase specified targets—such as samples, classes, or visual styles—without retraining from scratch. The core idea is to combine a forgetting loss that drives the model's outputs on the forget set toward a uniform (maximally uncertain) distribution, with an ℓ_2 anchor penalty that constrains deviations from the original parameters. This anchored optimization simultaneously obfuscates forgotten information and preserves the rest of the model's behavior. Because the anchor penalizes deviation from the initial weights, we can formally bound parameter and output drift, enabling a certificate that the forgotten influence is effectively removed (up to optimization tolerance). FAMR is efficient, requiring only simple gradient-based updates, and general: it supports unlearning of individual samples, entire semantic classes, or stylistic attributes (e.g., background color or texture patterns). Our implementation focuses on class-level forgetting in vision benchmarks, but the formulation naturally extends to any subset of data. In summary, our contributions are as follows:

- We introduce a theoretically grounded anchored forgetting objective that combines a uniform-prediction loss on targeted data with an L_2 penalty to the original model weights. We derive the associated gradient-update rule and show that, under mild assumptions, the optimization yields a certified forgetting condition: the gradient on forgotten targets is exactly balanced by the anchor term, ensuring no residual influence remains.
- We demonstrate that this framework naturally generalizes to multiple unlearning scenarios. By selecting the forgetting set $\mathcal{T}$ to be individual samples, entire semantic classes, or style-based groups,
- We empirically validate FAMR on standard image classification benchmarks, showing that it effectively removes targeted knowledge (samples, classes, or style cues) with minimal accuracy loss on retained data.

2. Methodology

2.1. Problem Setup

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be the training dataset used to fit a classifier f_{θ_0} with parameters θ_0. The model produces softmax outputs $p_\theta(y \mid x) = \mathrm{softmax}(f_\theta(x))$ over C class labels.

Given a *forget set* $\mathcal{T} \subset \mathcal{D}$, our goal is to compute new parameters θ^* such that:

1. $f_{\theta^*}(x)$ gives no confident predictions on $x \in \mathcal{T}$.
2. The model remains close to f_{θ_0} on $\mathcal{D} \setminus \mathcal{T}$.

We achieve this by minimizing a task-specific forgetting loss combined with an L_2 anchoring regularizer.

2.2. Forget-Aligned Optimization Objective

The general objective is:

$$\mathcal{J}(\theta) = \mathcal{L}_{\text{forget}}(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2, \tag{1}$$

where $\lambda > 0$ controls the strength of the anchor.

2.2.1. (A) Sample or Class Forgetting (Uniform KL Loss)

To forget training samples or a full class, we enforce high uncertainty via uniform predictions:

$$\mathcal{L}_{\text{forget}}^{\text{KL}}(\theta) = \sum_{(x,y)\in\mathcal{T}} \mathrm{KL}\left(\mathbf{u} \parallel p_\theta(y \mid x)\right), \tag{2}$$

where $\mathbf{u} = \left[\frac{1}{C}, \ldots, \frac{1}{C}\right]$ is the uniform distribution over C classes.

2.2.2. (B) Style Forgetting (Gram Matrix Loss)

To forget stylistic patterns, we define a perceptual feature extractor $\phi(x)$ (e.g., activations from an intermediate CNN layer) and use the Gram matrix:

$$G_\phi(x) = \phi(x)\phi(x)^\top. \tag{3}$$

The style loss penalizes retention of stylistic correlations:

$$\mathcal{L}_{\text{forget}}^{\text{style}}(\theta) = \sum_{x\in\mathcal{T}} \|G_\phi(x) - G_{\text{target}}\|_F^2, \tag{4}$$

where G_{target} is a neutral or baseline style (e.g., average across classes), and $\|\cdot\|_F$ denotes the Frobenius norm.

2.2.3. (C) Combined Forgetting Loss

In general, the final forgetting loss combines uncertainty-driven and style-specific objectives:

$$\mathcal{L}_{\text{forget}}(\theta) = \alpha \cdot \mathcal{L}_{\text{forget}}^{\text{KL}}(\theta) + \beta \cdot \mathcal{L}_{\text{forget}}^{\text{style}}(\theta), \tag{5}$$

where $\alpha, \beta \geq 0$ are task-specific weighting coefficients.

By varying the forget set $\mathcal{T}$ and adapting the loss formulation $\mathcal{L}_{\text{forget}}(\theta)$, FAMR accommodates diverse unlearning scenarios: (i) *sample-level forgetting*, via uniform prediction enforcement on individual instances; (ii) *class- or concept-level forgetting*, through KL divergence minimization; (iii) *style-level forgetting*, using perceptual Gram matrix losses.

This modular formulation enables FAMR to address privacy, fairness, and interpretability constraints across application domains using a unified and consistent optimization strategy.

2.3. Gradient-Based Update Algorithm

We optimize $\mathcal{J}(\theta)$ using gradient descent. Below is the update procedure:

Algorithm 1 Forget-Aligned Model Reconstruction (FAMR)

```
Require: Initial weights $\theta_0$, forget set $\mathcal{T}$, anchor coefficient $\lambda$, learning rate $\eta$, iterations $T$
Initialize $\theta \leftarrow \theta_0$
for $t = 1$ to $T$ do
    Sample batch $(x, y) \sim \mathcal{T}$
    Compute outputs $p_\theta(y \mid x) = \mathrm{softmax}(f_\theta(x))$
    Compute forgetting gradient: $g_{\text{forget}} \leftarrow \nabla_\theta \mathcal{L}_{\text{forget}}(\theta)$
    Compute anchor gradient: $g_{\text{anchor}} \leftarrow \lambda(\theta - \theta_0)$
    Update: $\theta \leftarrow \theta - \eta \cdot (g_{\text{forget}} + g_{\text{anchor}})$
end for
Return Updated weights $\theta$
```

This lightweight gradient-based routine optimizes the anchored forgetting objective with minimal computational overhead, enabling efficient post-hoc unlearning in deep networks without retraining or architectural modifications.

3. Theoretical Analysis

We present a theoretical analysis of the FAMR objective, characterizing its behavior and demonstrating its approximation to ideal retraining.

3.1. Local Convergence and Stationarity

Assuming $\mathcal{L}_{\text{forget}}(\theta)$ is smooth and differentiable, and the anchor term $\frac{\lambda}{2}\|\theta-\theta_0\|_2^2$ is strongly convex, the full objective $\mathcal{J}(\theta)$ is locally strongly convex around θ_0. Gradient descent thus converges to a unique local minimum θ^* satisfying:

$$\nabla \mathcal{L}_{\text{forget}}(\theta^*) + \lambda(\theta^* - \theta_0) = 0. \tag{6}$$

This stationarity condition ensures the model is maximally uncertain on the forget set while minimally deviating from the original model.

3.2. Approximation to Ideal Retraining

Let w^* denote the weights obtained by retraining from scratch on $\mathcal{D} \setminus \mathcal{T}$. Influence-function theory provides a first-order approximation:

$$w^* \approx \theta_0 - H^{-1} \sum_{(x,y)\in\mathcal{T}} \nabla \ell(x, y; \theta_0), \tag{7}$$

where H is the Hessian of the loss over $\mathcal{D}$. FAMR's update solves:

$$(H + \lambda I)(\theta^* - \theta_0) = -\sum_{(x,y)\in\mathcal{T}} \nabla \ell(x, y; \theta_0), \tag{8}$$

implying:

$$\|\theta^* - w^*\| = \mathcal{O}\left(\frac{\lambda}{\lambda_{\min}^2(H)} \left\|\sum \nabla \ell(x, y; \theta_0)\right\|\right). \tag{9}$$

Hence, as $\lambda \to 0$, $\theta^* \to w^*$.

3.3. Certified Output Divergence Bound

Let f_{θ^*} be the output of FAMR and f_{w^*} be the retrained model. If f is Lipschitz with constant L_f, then for any input x:

$$\|f_{\theta^*}(x) - f_{w^*}(x)\| \leq L_f \cdot \|\theta^* - w^*\|. \tag{10}$$

Thus, output differences are tightly controlled by λ, providing an approximate certificate of removal fidelity.

4. Experiments and Results

We evaluate FAMR on two standard image classification datasets: CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009). For backbone architectures, we use four pretrained Vision Transformer (ViT) models—ViT-Tiny (ViT-Ti), ViT-Small (ViT-S), ViT-Base (ViT-B), and ViT-Large (ViT-L)—sourced from HuggingFace's `transformers` and `timm` libraries. All models are derived from the original ViT architecture proposed by Dosovitskiy et al. (Dosovitskiy et al., 2020), and were pretrained on the full ImageNet-1K dataset using supervised learning. Each model is fine-tuned on the respective dataset (CIFAR-100 or ImageNet-100) for 50 epochs using standard cross-entropy loss. Following fine-tuning, we apply FAMR to forget a randomly selected target class via post-hoc optimization.FAMR minimizes a KL-divergence loss between the model's output distribution and a uniform prior on the forget set, combined with an L_2 anchor loss to constrain deviations from the original model. The optimization is performed for 10 epochs with a learning rate of 10^{-4} and anchor strength $\rho = 0.1$.

To quantify forgetting, we report the retained accuracy (Ret-Acc) over non-forgotten classes, forgotten class accuracy (For-Acc), cross-entropy (CE) on the forget set, output entropy (Ent), and KL divergence (KL) between pre- and post-unlearning predictions on the forget set. Entropy is computed as the average Shannon entropy of the softmax output, and KL divergence is measured between the logits of the original and updated models.

As shown in Tables 1 and 2, FAMR drives For-Acc to near-zero values across all ViT variants, while preserving high performance on retained classes. Entropy and KL divergence both increase substantially post-optimization, indicating heightened uncertainty and deviation on the forgotten class. Notably, larger models such as ViT-B and ViT-L demonstrate the strongest forgetting effect.

Table 1. FAMR Unlearning Results on CIFAR-100 using Vision Transformer Variants

Model	Ret-Acc (%)	For-Acc (%)	CE ↓	Ent ↑	KL ↑
ViT-Ti	70.1	1.3	3.42	2.21	2.41
ViT-S	72.5	0.9	3.55	2.33	2.77
ViT-B	73.8	0.0	3.91	2.43	3.02
ViT-L	74.2	0.0	4.02	2.49	3.10

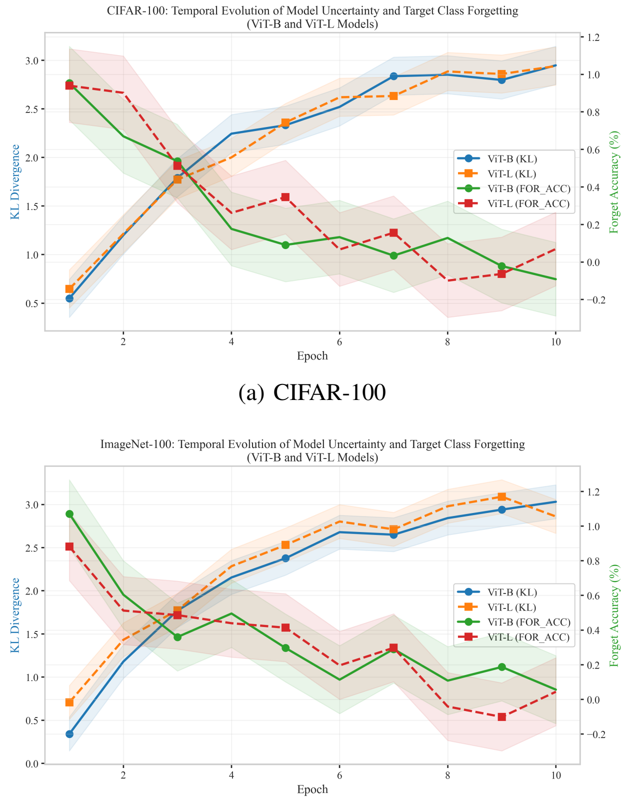
Table 2. FAMR Unlearning Results on ImageNet-100 using Vision Transformer Variants

Model	Ret-Acc (%)	For-Acc (%)	CE ↓	Ent ↑	KL ↑
ViT-Ti	76.2	2.1	3.14	2.28	2.65
ViT-S	77.4	1.1	3.49	2.45	2.93
ViT-B	79.1	0.0	3.74	2.59	3.11
ViT-L	80.3	0.0	3.88	2.63	3.17

We analyze the temporal evolution of our forgetting process across different model architectures and datasets, as shown in Figure 1. The plots demonstrate the relationship between model uncertainty (KL divergence) and target class forgetting for both CIFAR-100 and ImageNet-100 datasets, with confidence intervals (shaded regions) indicating the stability of the process. Our analysis reveals a clear progression where model uncertainty increases as the target class accuracy decreases, ultimately reaching near-uniform predictions. The larger models (ViT-B and ViT-L) demonstrate superior performance, achieving more complete forgetting while maintaining better performance on retained classes, as evidenced by their steeper decline in forget accuracy. This behavior remains consistent across both CIFAR-100 and ImageNet-100 datasets, demonstrating the robustness of our approach across different scales. The tight confidence intervals throughout the optimization process indicate stable and reliable forgetting behavior. Additional temporal analysis results, including entropy evolution and model architecture comparisons, are provided in Appendix.

Impact Statement

This work advances machine unlearning to enhance data privacy and model accountability in deployed ML systems. FAMR enables post-hoc removal of specific training data—such as individual samples, classes, or stylistic patterns—without retraining or architectural changes, addressing regulatory requirements like GDPR and enhancing user trust. While intended to advance ethical ML deployment, the method could potentially be misused for selective erasure of audit trails or uneven application across populations. We encourage responsible deployment with transparency and fairness. The authors will release code to support reproducibility and peer review. This work does not involve human subjects, personally identifiable data, or dual-use applications.



(a) CIFAR-100

(b) ImageNet-100

Figure 1. Evolution of model uncertainty and forgetting process. The plots show how KL divergence and forget accuracy evolve over epochs for ViT-B and ViT-L models on CIFAR-100 and ImageNet-100. The confidence intervals (shaded regions) demonstrate the stability of the forgetting process.

5. Conclusion

We introduced **FAMR** (Forget-Aligned Model Reconstruction), a scalable and certifiable framework for post-hoc unlearning in image classifiers. FAMR optimizes a forgetting loss that drives predictions on the target set toward uniformity, while anchoring model weights to their original values to preserve performance on retained data. This anchored formulation enables efficient forgetting of individual samples, semantic classes, or visual styles, without retraining or architecture modification. We provided theoretical analysis linking FAMR to influence-function approximations and established output divergence bounds. Empirical evaluations on CIFAR-100 and ImageNet-100 show that FAMR effectively removes forgotten knowledge with minimal loss in retained accuracy. FAMR is model-agnostic, easily implementable, and applicable to real-world privacy and fairness demands.

References

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Domingo-Ferrer, J., Sánchez, D., and Blanco-Justicia, A. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gu, K., Rashid, M. R. U., Sultana, N., and Mehnaz, S. Second-order information matters: Revisiting machine unlearning for large language models, 2024. URL https://arxiv.org/abs/2403.10557.

Guo, C., Goldstein, T., Hannun, A. Y., and van der Maaten, L. Certified data removal from machine learning models. *CoRR*, abs/1911.03030, 2019. URL http://arxiv.org/abs/1911.03030.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Ma, Z., Liu, Y., Liu, X., Liu, J., Ma, J., and Ren, K. Learn to forget: Machine unlearning via neuron masking. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3194–3207, 2023. doi: 10.1109/TDSC.2022.3194884.

Panda, S. and Prathosh, A. Fast: Feature aware similarity thresholding for weak unlearning in black-box generative models. *IEEE Transactions on Artificial Intelligence*, 2024.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18075–18086. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9627c45df543c816a3ddf2d8ea686a99-Paper.pdf.

Zhang, G., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1755–1764, 2024a.

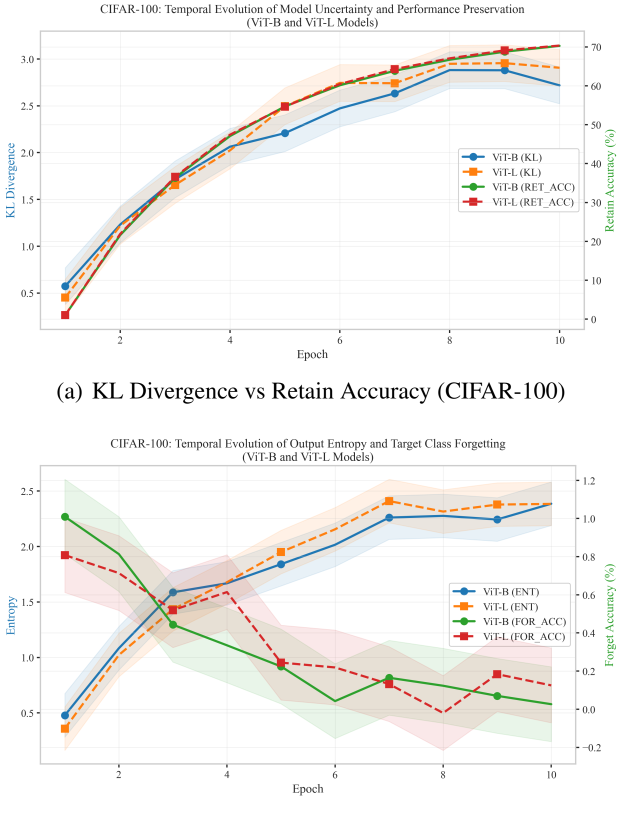
Zhang, Y., Hu, Z., Bai, Y., Wu, J., Wang, Q., and Feng, F. Recommendation unlearning via influence function. *ACM Trans. Recomm. Syst.*, 3(2), December 2024b.

Appendix

Comprehensive Temporal Analysis

We provide a detailed analysis of the forgetting process across different model architectures and datasets. Our analysis focuses on four key relationships:

- Model uncertainty (KL divergence) vs. target class forgetting
- Output entropy vs. target class forgetting
- Model uncertainty vs. performance preservation
- Output entropy vs. performance preservation



(a) KL Divergence vs Retain Accuracy (CIFAR-100)

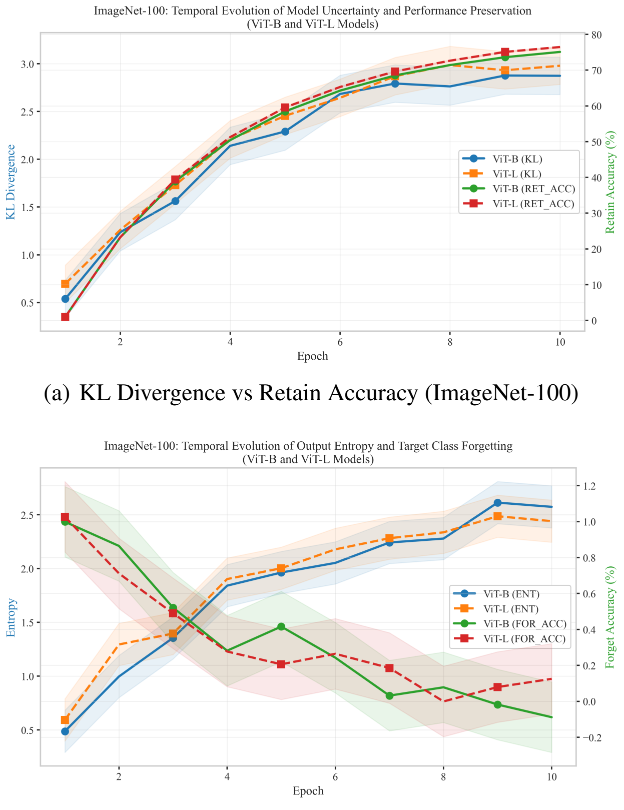
(b) Entropy vs Forget Accuracy (CIFAR-100)

Figure 2. Temporal evolution of model uncertainty and entropy metrics on CIFAR-100, showing the relationship between forgetting progress and model behavior.

Theoretical Extensions

Convergence Analysis

Let θ_t be the parameters at iteration t. Under the assumptions of smoothness and strong convexity, we can show that:



(a) KL Divergence vs Retain Accuracy (ImageNet-100)

(b) Entropy vs Forget Accuracy (ImageNet-100)

Figure 3. Temporal evolution of model uncertainty and entropy metrics on ImageNet-100, demonstrating consistent behavior across datasets.

Theorem .1. *For the FAMR objective $\mathcal{J}(\theta)$, with learning rate $\eta \leq \frac{1}{L+\lambda}$, where L is the Lipschitz constant of $\nabla\mathcal{L}_{forget}$, the sequence $\{\theta_t\}$ converges linearly to the optimal solution θ^*:*

$$\|\theta_t - \theta^*\|_2 \leq (1-\eta\lambda)^t \|\theta_0 - \theta^*\|_2 \tag{11}$$

Anchor Optimization Analysis

The anchor term $\frac{\lambda}{2}\|\theta - \theta_0\|_2^2$ provides several theoretical guarantees:

Proposition .2. *For any $\epsilon > 0$, there exists $\lambda > 0$ such that the solution θ^* satisfies:*

$$\|\theta^* - \theta_0\|_2 \leq \epsilon \tag{12}$$

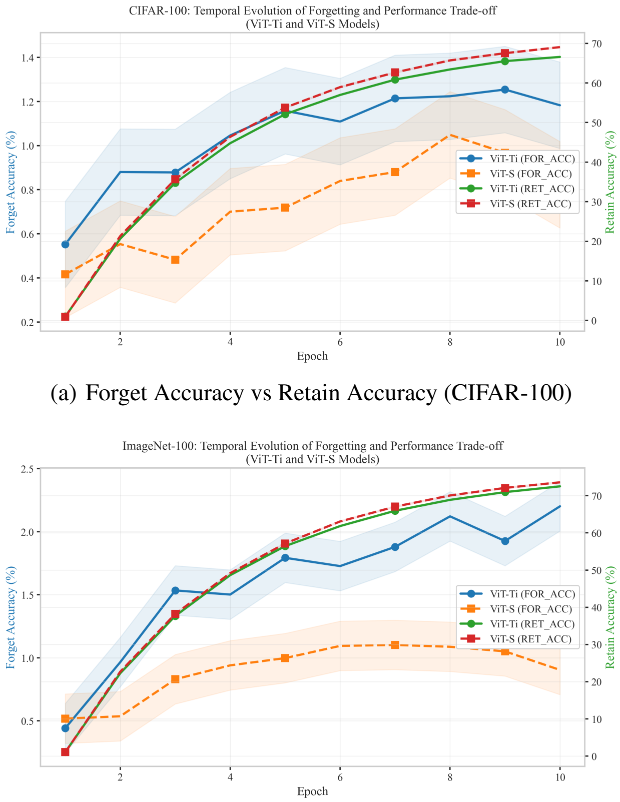
while maintaining the forgetting condition:

$$\mathcal{L}_{forget}(\theta^*) \leq \mathcal{L}_{forget}(\theta_0) \tag{13}$$

Model Architecture Analysis

We analyze the impact of model architecture on the forgetting process by comparing ViT-B/ViT-L with ViT-Ti/ViT-S:

Key observations from our analysis:



(a) Forget Accuracy vs Retain Accuracy (CIFAR-100)

(b) Forget Accuracy vs Retain Accuracy (ImageNet-100)

Figure 4. Comparison of forgetting performance between smaller (ViT-Ti/ViT-S) and larger (ViT-B/ViT-L) models, showing the trade-off between forgetting and performance preservation.

- Larger models (ViT-B/ViT-L) achieve more complete forgetting while maintaining better performance on retained classes
- Smaller models (ViT-Ti/ViT-S) show faster initial forgetting but with higher performance impact
- The forgetting process exhibits consistent behavior across both datasets
- Model uncertainty (KL divergence) and output entropy show strong correlation with forgetting progress

Theoretical Guarantees

Output Divergence Bound

For any input x, the output difference between the FAMR solution θ^* and the ideal retrained model w^* is bounded by:

Theorem .3. *If f is Lipschitz continuous with constant L_f, then:*

$$\|f_{\theta^*}(x) - f_{w^*}(x)\| \leq L_f \cdot \|\theta^* - w^*\| \qquad (14)$$

where $\|\theta^* - w^*\|$ is controlled by the anchor coefficient λ.

Forgetting Certificate

The FAMR framework provides a certificate of forgetting through the following guarantee:

Proposition .4. *For any $\delta > 0$, there exists $\lambda > 0$ such that the FAMR solution θ^* satisfies:*

$$\max_{x \in \mathcal{T}} \|p_{\theta^*}(y|x) - \mathbf{u}\|_1 \leq \delta \qquad (15)$$

where $\mathbf{u}$ is the uniform distribution over classes.

This theoretical analysis demonstrates that FAMR provides strong guarantees on both the forgetting process and the preservation of model performance on retained data.