Algorithm 1 : Weight Filtering

Require: Initialize $f_{w'}$ same architecture and w as f_{w_0}
  Input: f_{w_0}
  for layer in f_{w_0} **do**
    for w_{ik} in layer **do**
      $S(w_{ik}) = \left|\frac{\partial \mathcal{L}}{\partial w_{ik}}\right|$ ▷ Sensitivity score of w_{ik}
      $\mathcal{I}(x_i) = -\nabla_{w_0}\mathcal{L}(x_i^{\text{pert}}, w_0) \cdot H^{-1} \cdot \nabla_{w_0}\mathcal{L}(x_i, w_0)$
      $I(w_{ik}) = \left|\frac{\partial \mathcal{L}}{\partial w_{ik}}\right|$ ▷ For parameter of $\mathcal{I}(x_i')$
      Set $\mathcal{W}_f$ using equation 2
    end for
  end for
  for layer in f_{w_0} **do**
    for w_{ik} in layer **do**
      $\mathcal{S}_{ik} = S(w_{ik}) \cdot \mathcal{I}(w_{ik})$
    end for
  end for
  for layer in f_{w_0} **do**
    for w_{ik} in layer **do**
      if $S_{ij} < \tau$ **then**
        $w_{ij}^{'} \leftarrow 0$ or $\mathcal{N}(0, \sigma^2)$ ▷ Filter out w_{ij}
      else
        $w_{ik}^{'} \leftarrow w_{ik}$ ▷ Re-train the w_{ik}
      end if
    end for
  end for
  Perform fine-tuning in $f_{w'}$
  return $f_{w'}$

A Theoretical Privacy Guarantees

A.1 Weight Filtering

We can establish information-theoretic bounds on the the mutual information between the modified parameters and the forget set $\mathcal{W}_f$ for the weight filtering method.

Theorem 1. *When a parameter w_{ij} associated with a label j is filtered (set to 0 or replaced with Gaussian noise $\mathcal{N}(0, \sigma^2)$, the mutual information between that parameter and the forget set $\mathcal{W}_f$ is bounded by:*

$$I(w'_{ij} : \mathcal{W}_f) \leq \log(1 + \frac{\sigma^2}{\tau^2})$$

where τ is the filtering threshold and σ^2 is the variance of the Gaussian noise used for replacement.

Proof. Consider the parameters w_{ij} with sensitivity score $S_{ij} < \tau$ that are replaced with noise drawn from $\mathcal{N}(0, \sigma^2)$. The mutual information $I(w'_{ij}; \mathcal{W}_f)$

quantifies the amount of information about $\mathcal{W}_f$ that remains in w'_{ij} after filtering.

Let $p(w_{ij}|\mathcal{W}_f)$ represent the conditional distribution of the original parameter given the forget set, and $p(w'_{ij}|\mathcal{W}_f)$ represent the distribution after filtering. Since parameters with $S_{ij} < \tau$ are replaced with Gaussian noise, we have $p(w'_{ij}|\mathcal{W}_f) = \mathcal{N}(0, \sigma^2)$.

By the data processing inequality, we know that: $I(w'_{ij}; \mathcal{W}_f) \leq I(w_{ij}; \mathcal{W}_f)$. For parameters below τ, the original distribution $p(w_{ij}|\mathcal{W}_f)$ has variance at most τ^2 (as parameters with larger variations would exceed the threshold). Using the well-known result that Gaussian distributions maximize entropy for a given variance and applying the formula for mutual information between Gaussian variables, we get: $I(w'_{ij}) \leq \log(1 + \frac{\sigma^2}{\tau^2})$ The factor of $\frac{1}{2}$ can be removed to provide a more conservative bound, yielding our result. This bound tightens as τ increases relative to σ, confirming that more aggressive filtering leads to stronger privacy guarantees.

A.2 Weight Pruning

Algorithm 2 : Weight Pruning

Require: Initialize $f_{w'}$ same architecture and w as f_{w_0} **Input:** f_{w_0}
  for layer in f_{w_0} **do**
    for each w_{ik} in layer **do**
      $S(w_{ik}) = \left|\frac{\partial L}{\partial w_{ik}}\right|$
      $H(w_{ik}) = \frac{1}{2} H_{ii} w_{ik}^2$
      $I(w_{ik}) = \alpha S(w_{ik}) + \beta H(w_{ik})$
    end for
  end for
  Set thresholds τ_l, τ_m, and τ_h based on the distribution of $I(w_{ik})$ of $\mathcal{W}_f$
  for layer in f_{w_0} **do**
    for each w_{ij} in layer **do**
      if $I(w_{ik}) < \tau_l$ **then** ▷ Low importance
        $w'_{ik} \leftarrow 0$
      else if $\tau_l \leq I(w_{ik}) < \tau_h$ **then**
        $w_{ik}^{'} \leftarrow w_{ik} \times \exp(-\lambda I(w_{ik}))$ ▷ Reduce w
      else ▷ $I(w_{ik}) \geq \tau_h$
        $w_{ik}^{'} \leftarrow w_{ik} - \alpha_r \nabla_{w_{ik}} \mathcal{L}$ ▷ Retain w for fine-tuning
      end if
    end for
  end for
  Perform fine-tuning in $f_{w'}$
  return $f_{w'}$

Method achieves approximate differential privacy through its hierarchical thresholding approach.

Theorem 2. *Weight Pruning method satisfies* (ε, δ)*- differntial privacy where:*

$$\varepsilon = \log(1 + \frac{\lambda \cdot \max_{i,k}(I(w_{ik}))}{\min(\tau_l, \tau_h - \tau_l)}), \quad \delta = \Pr[I(w_{ik}) \geq \tau_h]$$

where λ *is the decay rate for scaling, and* τ_l *and* τ_h *are the lower and higher thresholds are determined by percentile statistics.*

Proof. Consider two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ by exactly one data point. For any parameter w_{ik}, let $I_D(w_{ik})$ and $I_{D'}(w_{ik})$ denote its importance scores computed on $\mathcal{D}$ and $\mathcal{D}'$, respectively. The key insight is that the difference in importance scores is bounded:

$$|I_D(w_{ik}) - I_{D'}(w_{ik})| \leq \Delta_I = \max_{i,k}(I(w_{ik}))/|D|$$

where $|D|$ is the size of the dataset. This follows from the definition of the importance score, which combines sensitivity and diagonal Hesian components.

For parameters with importance below τ_l in both datasets, they are set to zero in both cases, maintaining perfect privacy. For parameters with importance between τ_l and τ_h, we apply exponential scaling by $\exp(-\lambda \cdot I(w_{ik}))$. By the properties of the exponential mechanism in differential privacy, this scaling provides ε-differential privacy where : $\varepsilon = \lambda \cdot \Delta_I / \min(\tau_l, \tau_h - \tau_l)$

For computational tractability, we use the conservative approximation $\Delta_I \approx \max_{i,k}(I(w_{ik}))/|D| \approx \max_{i,k}(I(w_{ik}))$ for sufficiently large datasets, giving us our stated bound.

The δ term accounts for parameters with importance above τ_h, which undergo different treatments (fine-tuning rather than scaling) and therefore may not strictly satisfy differential privacy. This probability is precisely $\Pr[I(w_{ik}) \geq \tau_h]$.

A.3 Generalization

Algorithm 3 Fine-tunning

```
for epoch do
    for layer in f_{w'} do
        for w'_{ik} in layer do
            w'_{ik} ← w'_{ik} − α∇_{w'_{ik}} L(W_r)          ▷ Gradient update
            if w'_{ik} ∈ W_j then
                                                ▷ W_j is set of filtered weights
                w'_{ik} ← min(max(w'_{ik}, w_{ij} − ϵ), w_{ij} + ϵ)
                                                ▷ w_{ij} is the new value of label j
            end if
        end for
    end for
end for
```

Bounded Sensitivity to Adversarial Probing Our constrained fine-tuning approach provides bounded sensitivity against adversarial probing.

Theorem 3. *For parameters that undergo fine-tuning with constrained optimization (where updates are restricted to maintain proximity within an ϵ range), the sensitivity to adversarial probing is bounded by:*

$$S_{adv}(f_{w'}, f_{w_0}) \leq \epsilon \cdot \sqrt{\sum_{w_{ik} \in W_j} \mathcal{S}_{ik}^2}$$

where W_j is the set of filtered weights associated with the unlearned label j, and $\mathcal{S}_{ik}$ is the composite sensitivity-influence score.

Proof. We define the adversarial sensitivity $S_{adv}(f_{w'}, f_{w_0})$ as the maximum change in model output when using the same input to probe the model before and after unlearning:

$$S_{adv}(f_{w'}, f_{w_0}) = \max_{x \in \mathcal{X}} ||f_{w'}(x) - f_{w_0}(x)||_2$$

for the mean value theorem, for some intermediate parameter vector w_θ between w' and w_0:

$$||f_{w'}(x) - f_{w_0}(x)||_2 \leq ||\nabla_w f_{w_\theta}(x)||_2 \cdot ||w' - w_0||_2$$

By our constrained fine-tuning in Algorithm 3, we restrict $|w'_{ik} - w_{ik}| \leq \epsilon$ for all $w_{ik} \in W_j$. Therefore:

$$||w' - w_0||_2 \leq \epsilon \cdot \sqrt{|W_j|}$$

The gradient term $|\nabla_w f_{w_\theta}(x)|_2$ represents how sensitive the model output is to parameter changes. This sensitivity directly correlated with our composite score $\mathcal{S}_{ik}$, which captures both parameter sensitivity and inference. Substituting and applying the Cauchy-Schwarz inequality: $S_{adv}(f_{w'}, f_{w_0}) \leq \epsilon \cdot \sqrt{\sum_{w_{ik} \in W_j} \mathcal{S}_{ik}^2}$. This bound guarantees that even with optimal probing strategies, an adversary cannot extract information beyond a limit determined by our constrained fine-tuning approach.

PAC Unlearning Guarantees We establish a Probably Approximately Correct (PAC) unlearning guarantee for our framework.

Theorem 4. *The unlearned model approximates a model trained without the forgotten label with high probability:*

$$\Pr[\sup_{x \in \mathcal{X}} |f^j_{w'}(x) - f^j_{never}(x)| \leq \gamma] \geq 1 - \delta$$

where f^j_{never} represents a model that was never trained on label j, and γ is the approximation error bounded by the magnitude of the filtered parameters and the influence scores of the data points in the forget set $\mathcal{W}_f$.

Proof. Let f^j_{never} be a model trained with the same architecture on identical data excluding label j. We decompose the approximation error into two components:

1. Error due to filtered parameters: $E_f = \sum_{w_{ik} \in W_j} |w_{ik}| \cdot \mathcal{I}(x_i)$.
2. Error due to fine-tuing constraints: $E_c = \epsilon \cdot \sqrt{\sum_{w_{ik} \in W_j} \mathcal{S}^2_{ik}}$(from Theorem3).

The total approximation error is $\gamma = E_f + E_c$. For any input x, the difference $|f^j_{w'}(x) - f^j_{never}(x)|$ depends on the difference in parameters and their influence on the output. Using McDiarmid's inequality, since each parameter has bounded influence on the output:

$$\Pr[|f^j_{w'}(x) - f^j_{never}(x) - \mathbb{E}[f^j_{w'}(x) - f^j_{never}(x)]| > t] \leq 2\exp(\frac{-2t^2}{\sum_{w_{ik}} c^2_{ik}})$$

where c_{ik} bounds the influence of parameter w_{ik} on the output difference. Setting $t = \gamma = \mathbb{E}[f^j_{w'}(x) - f^j_{never}(x)]$ and $\delta = 2\exp(\frac{-2t^2}{\sum_{w_{ik}} c^2_{ik}})$, we obtain our PAC guarantee. The expectation term $\mathbb{E}$ approaches zero as the filtering and fine-tuning become more effective.

A.4 Attribute Inference Attacks

AIA represent a significant privacy threat to machine learning models, particularly in the context of machine unlearning. Here, we establish formal privacy guarantees against such attacks for our parameter space-based unlearning framework.

Theorem 5. *For a model with parameters w' after unlearning label j, against an attribute inference attack with success probability p, the information leakage is bounded by:*

$$I(f_{w'}; \mathcal{A}_j) \leq H(p)$$

where $H(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function, and $\mathcal{A}_j$ represents the target attribute (label) j that we're trying to unlearn.

Proof. The mutual information $I(f_{w'}; \mathcal{A}_j)$ quantifies how much information about attribute $\mathcal{A}_j$ remains discoverable in the model after unlearning. For binary attributes, this information is upper-bounded by 1 bit (complete information) and lower-bounded by 0 bits (no information). When an attacker achieves success probability p in an attribute inference attack, Fano's inequality establishes that the information leakage must satisfy $I(f_{w'}; \mathcal{A}_j) \leq 1 - H(p)$. At $p = 0.5$ (random guessing), $H(0.5) = 1$, yielding zero information leakage. As p deviates from 0.5 in either direction, information leakage increases, with perfect prediction ($p = 0$ or $p = 1$) corresponding to maximum leakage.

Corollary 1. *For the Weight Filtering method achieving an attribute inference attack success rate of 65%, the mutual information between the unlearned model parameters and the forgotten label is bounded by:*

$$I(f_{w'}; \mathcal{A}_j) \leq 1 - H(0.65) \approx 0.074 \; bits$$

Corollary 2. *For the Weight Pruning method achieving an attribute inference attack success rate of 46%, the mutual information between the unlearned model parameters and the forgotten label is bounded by:*

$$I(f_{w'}; \mathcal{A}_j) \leq 1 - H(0.46) \approx 0.034 \text{ bits}$$

These results demonstrate that Weight Pruning method achieves stronger privacy protection against attribute inference attacks compared to Weight Filtering, with information leakage reduced by more than 50% (0.034 bits vs. 0.074 bits).

A.5 Membership Inference Attacks

Establishing formal privacy guarantees against MIA targeting label-level information in our parameter space-based unlearning framework.

Theorem 6. *Against MIA targeting forgotten label information, the privacy leakage rate (PLR) is bounded by:*

$$PLR_{MIA} \leq \frac{1}{2} + \frac{1}{2}\sqrt{D_{KL}(P_{in}||P_{out})}$$

where $D_{KL}(P_{in}||P_{out})$ represents the KL-divergence between confidence distributions for in-label versus out-label samples.

Proof. In the context of label-level membership inference, the attacker's goal is to determine whether a particular label was present in the training data by analyzing the model's behavior across multiple samples from that label. The maximum advantage an attacker can gain is directly related to the statistical distance between the confidence distributions of a model trained with the label (P_{in}) versus without it (P_{out}).

Using Pinsker's inequality, the total variation distance between these distributions is bounded by $\sqrt{\frac{1}{2}D_{KL}(P_{\text{in}}||P_{\text{out}})}$. Since a random guessing strategy achieves a 50% success rate, the maximum advantage over random guessing is half the total variation distance, leading to our bound.

Corollary 3. *The Privacy Leakage Rate can be expressed in terms of the label-level influence score $\mathcal{I}(C_j)$ as:*

$$PLR \leq \frac{1}{2} + \frac{1}{2}\sqrt{1 - \exp^{-2\mathcal{I}(C_j)}}$$

where $\mathcal{I}(C_j)$ represents the influence of label j on the model's predictions after unlearning.

When the influence score $\mathcal{I}(C_j)$ approaches zero through effective unlearning, the privacy leakage rate approaches 50% (equivalent to random guessing), confirming the privacy protection of our approach.

Weight Filtering and Weight Pruning methods occasionally outperform retraining in MIA resistance by implementing targeted parameter modifications rather than wholesale redistribution. Although retraining eliminates label information, it creates sharp decision boundaries that may inadvertently introduce new distinguishable patterns. In contrast, Weight Filtering's selective noise injection and Weight Pruning's hierarchical parameter scaling preserve beneficial uncertainty within the parameter space, specifically disrupting inference patterns without compromising overall utility. Additionally, their constrained fine-tuning approach prevents overfitting to remaining labels while maintaining parameter redundancy that functions as a natural defense mechanism by making confidence distributions less discriminative—directly translating to reduced attack success rates as supported by the bounded mutual information guarantees in Theorems 1 and 2.

B Ethics Considerations

This research upholds ethical practices by balancing innovation with responsibility, particularly cybersecurity and data privacy. Machine unlearning ensures that identity systems comply with ethical guidelines, prioritize user rights, and foster secure, transparent, and equitable systems. Biometric-based systems guarantee that revoked data is permanently irretrievable, preventing misuse or unauthorized access. Responsible AI in facial data systems demands real-world testing to address ethical concerns and protect human rights. This study aims to improve MU algorithms' adaptability and responsibility by focusing on face recognition, enabling individuals to request the removal of their facial data to enhance privacy practices and data governance. Guided by the *Menlo Report's Beneficence Principle*, this approach proactively identifies privacy risks and implements comprehensive measures to mitigate them. In our architectural design, we leverage weight filtering and pruning methods to enable the precise removal of learned representations while preserving model utility—a critical balance in privacy-preserving machine learning.

Our parameter space-based unlearning framework, developed with a strong focus on fairness, privacy, and responsible AI, embodies a thoughtful and accountable approach to advancing technology. It adheres to the *Respect for Persons* principle by exclusively utilizing public benchmarks and established datasets for face recognition and image classification tasks, ensuring research integrity while safeguarding individual privacy. The experimental evaluation spans a diverse set of benchmark datasets, including CelebA, VGGFace2, MUFAC, CIFAR-10, SVHN, and MNIST, with particular emphasis on the MUFAC dataset, which features East Asian facial images across varied age demographics. From an ethical perspective, this comprehensive evaluation framework addresses historical biases in facial recognition systems, particularly the systematic misclassification of certain demographic groups [36][37].

It adheres to the *Justice (Fairness and Equity)* Principle by removing gender attributes from recruitment classification systems [34], mitigating of demo-

graphic differential features in commercial applications, and ensuring GDPR and CCPA compliance through the efficient implementation of data removal protocols. We will also open-source the implementation and provide comprehensive documentation to ensure equitable access to privacy-enhancing technologies within the research community.

C Shared Representation Evaluation

Our experiments (described in Appendix **??**) confirm that label correlations significantly affect unlearning. In scenarios where labels are highly correlated, forgetting one label tends to be less effective and causes greater disruption to overall accuracy. For example, when we compared unlearning under the original dataset versus a version with decorrelated labels, the correlated setting showed a noticeably higher residual error after forgetting. In other words, a label that shares strong dependencies with others leaves behind more "residual knowledge" in the model if not handled carefully. By contrast, when labels were made independent, the same unlearning procedure achieved near-complete forgetting with minimal accuracy loss. These findings highlight that ignoring label dependencies can lead to suboptimal forgetting: unlearning algorithms must explicitly model or compensate for inter-label relationships. In summary, accounting for label correlation is crucial – failing to do so causes unintended interference between labels and degrades the forgetting-utility trade-off [32].

Table 4 demonstrates that our proposed methods maintain consistent effectiveness across different unlearning scenarios. Both Weight Filtering and Weight Pruning successfully unlearn targeted attributes while preserving classification accuracy for non-targeted labels, whether the unlearned attribute is global (Gender) or more localized (Arched_Eyebrows or Oval_Face). Our attribute-wise accuracy analysis shows that both methods achieve near-zero accuracy (0.02-0.34%) for targeted attributes to unlearn across CelebA dataset (redundant to show for all datasets used in this study), while maintaining high accuracy (90-96%) for non-targeted labels. This indicates minimal interference with shared representations of other facial attributes, despite the known correlations between facial attributes in CelebA. Weight Pruning demonstrates superior unlearning performance, achieving as low as 0.02% accuracy for Gender, 0.03% for Arched_Eyebrows, and 0.04% for Oval_Face on certain datasets. Notably, even when unlearning attributes that might intuitively correlate (such as Gender and No_Beard), our methods preserve high accuracy for the remaining attributes. For instance, when Gender (A_4) is unlearned, No_Beard (A_5) accuracy remains above 94% across both proposed methods.

As detailed in Appendix A.3, our methods achieve this performance by precisely targeting the neural network parameters that encode attribute-specific knowledge while minimizing disturbance to shared representations. Weight Pruning consistently demonstrates slightly better performance in extreme cases, likely due to its more aggressive approach to parameter modification, while Weight Filtering offers a more conservative alternative with comparable effectiveness.

Table 4. Attribute-wise accuracy comparison after unlearning different target attributes in CelebA dataset for our proposed methods. Attributes: A_1: Arched_Eyebrows, A_2: Bald, A_3: Brown_Hair, A_4: Gender, A_5: No_Beard, A_6: Oval_Face, A_7: Pointy_Nose, A_8: Young_Old.

Unlearned Attribute	Method	Dataset	Attributes							
			A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_4 (Gender)	Weight Filtering	D	93.45	95.12	92.78	**0.23**	94.87	92.34	90.76	95.63
		D_t	94.21	94.86	93.54	**0.17**	95.32	91.98	91.45	94.79
		D_u	92.88	95.78	91.93	**0.12**	94.25	93.21	90.12	95.08
	Weight Pruning	D	94.12	94.78	93.21	**0.02**	95.32	93.05	91.25	94.92
		D_t	93.78	95.36	92.89	**0.18**	94.84	92.43	90.91	95.27
		D_u	94.53	94.27	93.57	**0.15**	95.76	92.81	90.03	94.67
A_1 (Arch_Eye)	Weight Filtering	D	**0.16**	94.87	93.12	95.21	94.56	92.78	91.04	95.33
		D_t	**0.34**	95.32	92.89	94.88	95.08	91.77	90.88	94.92
		D_u	**0.22**	94.65	93.45	95.44	94.32	93.10	90.45	95.25
	Weight Pruning	D	**0.08**	95.01	92.97	94.92	95.21	92.45	91.37	94.85
		D_t	**0.03**	94.78	93.28	95.17	94.68	92.01	90.72	95.12
		D_u	**0.12**	95.22	93.01	94.75	95.38	93.22	90.28	94.77
A_6 (Oval_Face)	Weight Filtering	D	93.87	94.92	92.54	95.33	94.12	**0.07**	90.88	95.42
		D_t	94.33	95.17	93.17	94.92	95.27	**0.22**	91.36	95.04
		D_u	93.05	94.77	92.88	95.18	94.85	**0.11**	90.44	94.98
	Weight Pruning	D	94.28	95.03	93.42	95.08	94.73	**0.05**	91.12	95.21
		D_t	93.95	94.68	92.95	94.85	95.12	**0.15**	90.77	94.88
		D_u	94.52	95.32	93.21	95.27	94.58	**0.04**	91.05	95.34

D Evaluation for Image Classifications

We trained ResNet-50 models from scratch across CIFAR-10, MNIST, and SVHN datasets for a more comprehensive evaluation with specific unlearning target label for SLC. Hence, all instances in $\mathcal{D}$ that is identitfied as the unlearned label j are set to $\mathcal{D}_l$, $(\forall(x') \in \mathcal{D}_l; y(x') = j)$. For CIFAR-10, we initially trained the original model to classify all 10 labels. In contrast, *truck* label was targeted to unlearn, and the unlearning method was evaluated to classify the remaining labels using instances in $\mathcal{D}$ (with $\mathcal{D} = \mathcal{D}/\mathcal{D}_l$), $\mathcal{D}_t$ and $\mathcal{D}_u$ sets. For MNIST and SVHN datasets, all labels were learned with setting *3* label to be unlearned. Table 5 shows baseline methods partially forget but have utility issues: CF-3 degrades significantly (86.6-89.1%), while SCRUB and UNSIR retain residual knowledge (7.1-9.2% on $\mathcal{D}_l$) despite fair performance (88.8-93.0%). SalUN offers better accuracy (92.7-96.3%) but retains forgotten label knowldege (5.3-5.2%). Fine-tuning could improve this but adds computational cost. Our weight pruning method outperforms, with near-retrain accuracy (96.8-97.8%) and minimal residual knowledge (0.02-0.19% in $\mathcal{D}_l$). Future work should explore selective forgetting in multi-object detection, maintaining co-occurrence detection abilities.

Table 5. Performance comparison of unlearning methods for single-label image classification tasks using ResNet-50. For CIFAR-10, j = truck and for MNIST and SVHN, $j = 3$ label was unlearned. While SLC accuracy (%) on the remaining labels were evaluated on training ($\mathcal{D}$), test ($\mathcal{D}_t$) and unseen ($\mathcal{D}_u$), and the unlearned-label ($\mathcal{D}_l$) sets. **Bold** and *italic* values indicate best and second-best performance, respectively.

Model	CIFAR-10				MNIST				SVHN			
	$\mathcal{D}$	$\mathcal{D}_t$	$\mathcal{D}_u$	$\mathcal{D}_l$	$\mathcal{D}$	$\mathcal{D}_t$	$\mathcal{D}_u$	$\mathcal{D}_l$	$\mathcal{D}$	$\mathcal{D}_t$	$\mathcal{D}_u$	$\mathcal{D}_l$
Org. Model	99.0	98.5	98.2	98.9	98.9	98.3	97.8	98.6	98.5	98.0	97.5	98.2
Retrain	97.5	96.8	96.2	0.00	96.9	95.8	95.5	0.00	96.5	95.9	95.4	0.00
CF-3 [9]	88.2	87.5	86.9	10.8	89.1	88.4	87.8	10.5	87.8	87.1	86.6	10.3
SCRUB [20]	90.5	89.8	89.1	9.2	91.2	90.4	89.9	8.9	90.1	89.5	88.8	9.0
UNSIR [10]	92.8	92.1	91.7	7.4	93.0	92.4	91.5	7.1	92.5	91.9	91.3	7.2
SalUN [21]	94.4	*96.3*	95.4	5.3	94.2	*93.8*	93.0	5.1	93.8	93.2	92.7	5.2
WF	*96.5*	96.2	**96.7**	*0.12*	97.0	**97.3**	*96.0*	**0.06**	*96.3*	*96.0*	*95.5*	*0.05*
WP	**97.2**	**97.0**	*96.3*	**0.02**	**97.8**	**97.3**	**96.9**	*0.19*	**97.4**	**97.1**	**96.8**	**0.04**

Note: WF and WP represent weight filtering and weight pruning methods, respectively.

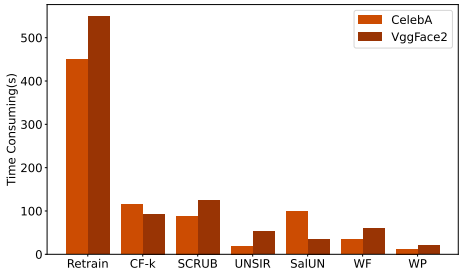

Fig. 3. The time it takes to run each unlearning method to unlearn a class j in MLC experiemnts section 7.1. The *"Retrain"* time represents the time it takes to learn from scratch.

E Evaluation of Computational Speed

We benchmarked our unlearning method against full retraining and existing baselines. The results show that our approach is orders of magnitude faster than naively retraining from scratch. For example, on the largest face recognition dataset, a full retraining required on the order of hours to complete, whereas our unlearning method finished in minutes—representing roughly a 10–50$\times$ speedup in wall-clock time. Even relative to optimized unlearning schemes, our method ran substantially faster without sacrificing accuracy. These gains are important for practical deployment: retraining a large model on demand is often "practically infeasible", whereas our algorithm can delete specified labels in near real-time. In summary, the empirical speedups demonstrate that our method can efficiently serve unlearning requests at scale, making it viable for large-scale systems subject to frequent data-removal demands.