FINU: Fisher-Informed Noise Injection for Efficient Zero-Shot Unlearning

Varun Sampath Kumar, Esmaeil S. Nadimi and Vinay Chakravarthi Gogineni
Applied AI and Data Science Unit, The Maersk Mc-Kinney Moller Institute
University of Southern Denmark, Odense, Denmark
{vaku, esi, vigo}@mmmi.sdu.dk

***Abstract*—Machine unlearning aims to remove the influence of specific training data from a deployed model in order to satisfy privacy and regulatory requirements. While exact unlearning methods such as retraining provide strong guarantees, they are computationally expensive and impractical at scale. Recent approximate unlearning approaches improve efficiency but typically rely on access to retained training data, limiting their applicability in privacy-constrained settings. This paper studies the more challenging *zero-shot* unlearning setting, where only the trained model and the data to be forgotten are available. We propose FINU, a Fisher-guided noise injection-based zero-shot unlearning framework that induces selective forgetting without using the retain dataset. FINU is motivated by the hierarchical structure of deep neural networks (DNNs) and employs an adaptive, layer-wise masking strategy based on Fisher Information to identify parameters most influential for the forget set. Controlled, learnable noise is then injected into the selected parameters to maximize the loss on forget samples, effectively removing their influence while preserving generalizable knowledge. We evaluate FINU across class-level, subclass-level, and sample-level unlearning scenarios on CIFAR-100, CIFARSuper20, and ImageNet-1k using prominent DNNs.**

***Index Terms*—Right to be forgotten, machine unlearning, data regulations**

I. Introduction

As modern deep learning models are trained on increasingly large and diverse datasets, concerns around data privacy, security, and regulatory compliance have become paramount. Once a model is trained, the influence of individual data points becomes implicitly encoded within its parameters, making simple deletion of samples from the underlying dataset insufficient to remove their effect [1]. Consequently, regulations such as GDPR [3] and PIPEDA [5] mandate mechanisms that enable the removal of specific data and its associated influence from deployed models, commonly referred to as the "right to be forgotten."

Machine unlearning [2], [20] is a promising solution to comply with these requirements by enabling models to remove the influence of a specified forget set, defined as the subset of training data whose contribution must be removed, without retraining from scratch, while preserving the performance of the retain set, which consists of the remaining training data. Existing unlearning methods can be broadly categorized into *exact* and *approximate* approaches. Exact methods, such as full retraining or SISA [6], guarantee complete removal of the forget set but are often computationally prohibitive for large-scale models. Approximate methods improve computational efficiency, but introduce a challenging trade-off between forgetting effectiveness, model utility, and computational cost. Most of the approaches rely on the use of the retain set to preserve utility, which introduces additional computational complexity, privacy risks, and limits applicability in scenarios where retained data are unavailable due to storage constraints, access restrictions, or oversight during model development.

Motivated by these limitations, recent works have introduced *zero-shot (ZS)* unlearning [7], [8], [12], where unlearning must be performed without using the retain dataset. In this setting, only the trained model and the forget data are available during unlearning. While more realistic in deployment scenarios, ZS unlearning is substantially more challenging, as the absence of retained data removes an explicit mechanism for safeguarding generalization performance. As a result, ZS unlearning methods must carefully balance effective forgetting with minimal disruption to the model.

Recent studies have shown that *localized unlearning*, where the unlearning algorithm operates on only a subset of model parameters, can offer a favorable balance between forgetting effectiveness, model utility, and computational efficiency [11], [21]. Motivated by these findings, an effective strategy for zero-shot unlearning is to restrict updates to carefully selected parameters rather than perturbing the entire network.

In this work, we study zero-shot unlearning from a perspective grounded in the parameterization and hierarchical structure of deep neural networks [19]. A key observation motivating our approach is that model parameters do not contribute equally to memorization and generalization: early layers tend to encode general-purpose features transferable across tasks, while deeper layers progressively specialize to task- and data-specific patterns. This hierarchical organization suggests that effective unlearning should selectively target the parameters most influential to the forget set while minimally perturbing general representations.

Leveraging these insights, we propose **FINU**, a zero-shot unlearning framework that removes forget-set-specific knowledge directly in parameter space without using the retain dataset. Our main contributions are summarized as follows:

- We propose FINU, a zero-shot unlearning framework that operates without the retain dataset and selectively re-

moves forget-set-specific knowledge via adaptive, Fisher-guided parameter masking and learnable noise injection.
- We demonstrate the effectiveness of FINU across class-level, subclass-level, and sample-level unlearning scenarios on multiple benchmarks and architectures.
- We provide qualitative evidence of effective forgetting through class activation map visualizations, illustrating the removal of forget-set-specific features.

Code Availability: The code for the proposed method is available at https://github.com/VarunSampath23/FINU-Unlearning.

II. Related Work

Most state-of-the-art machine unlearning methods rely on access to all or part of the original training data that is not requested for removal, commonly referred to as the *retain set*, and therefore do not satisfy the zero-shot constraint. For example, Bad Teacher unlearning [14] and SCRUB [15] employ student–teacher frameworks that explicitly leverage retain data during unlearning. Selective Synaptic Dampening (SSD) [13] uses Fisher information to identify important parameters, but requires both forget and retain data to estimate relative importance and applies a global dampening factor that is sensitive to hyperparameter tuning.

Early zero-shot unlearning approaches include Random labeling [17] where the labels of forget samples are randomized followed by fine-tuning. Zero-shot machine unlearning (EMMN) [12] introduces error-maximizing noise in the input space to degrade performance on forget samples, but this often leads to significant utility loss and does not directly target model parameters. Boundary-based methods [7] modify decision boundaries using adversarial label shifts, but either incur high computational cost (boundary shrinking) or achieve weaker forgetting performance (boundary expanding). More recently, just-in-time (JiT) unlearning [8] suppresses gradients in a local neighborhood of forget samples, demonstrating the effectiveness of localized, data-free updates.

FINU differs from prior work along three key dimensions. First, unlike SSD, it operates strictly in a zero-shot setting and avoids relative importance estimation using retain data. Second, instead of global dampening, FINU introduces an *adaptive, layer-wise masking* strategy guided by Fisher information, shifting from uniform parameter suppression to structured, selective perturbation. Third, unlike input- or gradient-based zero-shot methods (e.g., EMMN and JiT), FINU directly targets model parameters using an explicit importance measure, enabling more precise removal of forget-specific information while preserving general representations.

III. FINU: Fisher-Informed Noise Injection

Let f_θ denote a neural network with parameters θ trained on a dataset $\mathcal{D}$. Given a forget set $\mathcal{D}_f \subset \mathcal{D}$ requested for removal, the goal of machine unlearning is to obtain updated parameters θ' such that the influence of $\mathcal{D}_f$ is removed, while maintaining performance on the retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$. In the zero-shot unlearning setting considered in this work, $\mathcal{D}_r$ is unavailable during unlearning. Only the trained model f_θ and the forget set $\mathcal{D}_f$ are accessible. This setting is particularly challenging, as no explicit mechanism exists to directly preserve performance on the retain data.

A. Fisher-Guided Adaptive Parameter Masking

Deep neural networks exhibit a hierarchical structure in which earlier layers capture general, transferable features, while deeper layers encode increasingly task- and data-specific patterns. Consequently, parameters in deeper layers are more likely to store information specific to the forget set, whereas perturbations to early layers can significantly degrade overall model utility.

Motivated by this observation, FINU selectively targets parameters that are most influential for the forget set while preserving those associated with general representations. To identify such parameters, we estimate an empirical Fisher information measure with respect to $\mathcal{D}_f$. Specifically, for each parameter θ_i, we compute

$$\mathcal{I}_i = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\left(\frac{\partial \ell(f_\theta(x), y)}{\partial \theta_i}\right)^2\right],$$

where $\ell(\cdot)$ denotes the cross-entropy loss.

The Fisher information captures the sensitivity of model parameters to the forget data, with higher values indicating greater influence on predictions over $\mathcal{D}_f$.

FINU then applies an adaptive layer-wise masking strategy. For a network with L layers, we define a layer-specific masking ratio k_ℓ as

$$k_\ell = k_{\min} + \frac{\ell - 1}{L - 1}\left(k_{\max} - k_{\min}\right), \quad \ell \in \{1, \ldots, L\},$$

where $k_{\min}$ and $k_{\max}$ denote the minimum and maximum masking percentages, respectively. This assigns a smaller masking ratio to early layers and progressively increases it toward deeper layers.

For each layer ℓ, the top-k_ℓ% parameters are selected based on Fisher scores, ensuring that perturbations are concentrated in layers more likely to encode forget-specific information. Bias and normalization parameters are excluded from masking.

The resulting binary mask $m \in \{0, 1\}^{|\theta|}$ determines which parameters are eligible for perturbation during unlearning.

B. Learned Noise Injection for Forgetting

Given the adaptive mask m, FINU introduces a learnable noise parameters δ with the same dimensionality as θ. The perturbed parameters are defined as

$$\theta' = \theta + m \odot \delta,$$

where $\odot$ denotes Hadamard product (element-wise multiplication).

The noise parameters δ are optimized to degrade model performance on the forget set by maximizing the classification loss such as cross entropy loss $\mathcal{L}_{CE}$:

$$\max_{\delta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\mathcal{L}_{CE}(f_{\theta'}(x), y)\right] \; - \; \lambda \|\delta\|_2^2,$$

Algorithm 1 FINU: Fisher-guided Noise Injection based Zero-Shot Unlearning

Require: Trained model f_θ, forget set $\mathcal{D}_f$, epochs T, learning rate η, regularization λ
Ensure: Unlearned model $f_{\theta'}$

```
Initialize Fisher scores I ← 0
for each (x, y) ∈ D_f do
    I ← I + (∇_θ ℓ(f_θ(x), y))^2
end for
Generate adaptive layer-wise mask m from I
Initialize noise δ ← 0
for t = 1 to T do
    for each (x, y) ∈ D_f do
        θ' ← θ + m ⊙ δ
        L ← −L_CE(f_θ'(x), y) + λ‖δ‖_2^2
        Update δ ← δ − η∇_δ L
    end for
end for
θ ← θ + m ⊙ δ
return f_θ
```

where λ controls the magnitude of the perturbation and prevents excessive deviation from the original model. In our experiments, $\lambda = 1$ was found to be a stable choice for this hyperparameter across different settings.

During this optimization, the original model parameters θ remain frozen, and only δ is updated. This ensures that forgetting is induced solely through structured perturbations rather than retraining. After optimization, the learned noise is permanently applied to the model parameters to obtain the unlearned model.

IV. Experiments

A. Experiment Setup

Dataset – For our study, we leverage the CIFAR-100, CIFAR-Superclass-20 (CIFARSuper20), and ImageNet-1k datasets to evaluate the effectiveness and scalability of the proposed unlearning method across varying data complexities.

- **CIFAR-100** [18] is a widely used benchmark dataset consisting of 60,000 color images of resolution 32×32, evenly distributed across 100 fine-grained classes, with 500 training and 100 test images per class. This dataset enables evaluation of class-level unlearning performance in a controlled and compact setting.
- **CIFARSuper20** [14] is derived from CIFAR-100 by grouping its 100 fine-grained classes into 20 coarse-grained superclasses, each containing five semantically related subclasses. This setup allows us to study subclass-level unlearning, where the objective is to forget a specific subclass while preserving performance on the remaining correlated subclasses within the same superclass.
- **ImageNet-1k** [16] is a large-scale visual recognition dataset containing approximately 1.28 million training images and 50,000 validation images across 1,000 object categories. We use ImageNet-1k to assess the scalability of the proposed approach and its effectiveness in realistic, high-dimensional, and large-scale unlearning scenarios.

Model Architectures – We evaluate the proposed unlearning method across multiple neural network architectures to demonstrate its robustness and architecture-agnostic behavior. Specifically, we consider ResNet-18 [10], and ViT-B/16 [9], covering standard residual networks, and transformer-based architectures.

For all architectures, we initialize the models using weights pretrained on ImageNet-1k. For CIFAR-100 and CIFARSuper20, we fine-tune the pretrained models on the respective target datasets for 10 epochs using the Adam optimizer with a learning rate of 0.001, employing a ReduceLROnPlateau learning rate scheduler. For ImageNet-1k unlearning evaluation, we directly use the pretrained models without additional fine-tuning prior to performing unlearning.

Baselines – We evaluate the proposed method against several representative unlearning baselines. Retraining is included as the gold standard, where the model is trained from scratch using only the retain dataset after removing the forget data. SSD is also included as a strong baseline that leverages both the retain and forget datasets during unlearning. To ensure fair comparison in the retain-data-free (zero-shot) setting, we further include Random Labeling, Boundary Shrinking (BDSH), Just-In-Time Unlearning (JiT), and our proposed method (FINU), all of which operate using only the forget dataset during the unlearning phase.

For all methods, hyperparameters are tuned to achieve a reasonable trade-off between forgetting efficacy and retained model utility. Although access to both retain and forget data can enable improved Pareto trade-offs between forgetting efficacy and model utility, such access is not strictly necessary and feasible in many situations. Accordingly, we focus on the retain-data-free setting, where methods rely solely on the forget data during unlearning.

Evaluation Metrics – We evaluate FINU unlearning using a combination of accuracy-based, privacy-oriented, and efficiency metrics. We report black-box membership inference attack (MIA) success rates and wall-clock unlearning time (in seconds). For class-level and subclass-level unlearning, we measure retain accuracy ($\text{Acc}_{\mathcal{D}_r}$) on the test set of remaining classes and forget accuracy ($\text{Acc}_{\mathcal{D}_f}$) on the test set of the unlearned class/subclass. For sample-level unlearning, we additionally report test accuracy ($\text{Acc}_{\mathcal{D}_t}$) on the full remaining data to assess the trade-off between selective forgetting and overall generalization. Together, these metrics provide a comprehensive assessment of unlearning effectiveness, model utility preservation, and privacy protection.

In this section, we analyze the performance of the proposed FINU method across three unlearning settings: single-class unlearning, subclass unlearning, and random sample unlearning. The primary objective of these experiments is to evaluate the effectiveness of FINU in removing information associated with the specified forget set while preserving the overall performance of the model on the remaining data. All

TABLE I: Full class unlearning on ImageNet-1k with ViT-B/16.

Method	ZS	$\text{Acc}_{\mathcal{D}_r}\uparrow$	$\text{Acc}_{\mathcal{D}_f}\downarrow$	Time↓
Original	×	80.96	94.0	
SSD [13]	×	80.16	0.0	3368
Random Label [17]	✓	79.69	86.0	62
BDSH [7]	✓	80.87	70.0	130
JiT [8]	✓	72.78	18.0	45
FINU	✓	79.68	0.0	37

experimental results reported in this paper are averaged over three independent runs.

B. Full Class Unlearning

We consider class-level unlearning, where all training samples belonging to a target class k are removed. Let the training dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. The forget set is defined as

$$\mathcal{D}_f^{(k)} = \{(x_i, y_i) \in \mathcal{D} \mid y_i = k\}, \tag{1}$$

and the retain set is $\mathcal{D}_r^{(k)} = \mathcal{D} \setminus \mathcal{D}_f^{(k)}$. The objective is to eliminate the influence of $\mathcal{D}_f^{(k)}$ while preserving performance on $\mathcal{D}_r^{(k)}$.

Results. Table I reports the single-class unlearning results on ImageNet-1k using ViT-B/16. SSD achieve near-zero forget accuracy, confirming their effectiveness in removing class-specific information; however, SSD incurs a substantial computational cost, requiring over 3,000 seconds for unlearning.

Among retain-data-free baselines, Random Labeling and BDSH partially reduce forget accuracy but fail to completely eliminate the influence of the forget class, while also exhibiting varying degrees of retain accuracy degradation. JiT achieves stronger forgetting but at the expense of a significant drop in retain accuracy, indicating over-forgetting.

In contrast, the proposed FINU method achieves near-zero forget accuracy while maintaining competitive retain accuracy, closely matching retain-data-based methods, but with significantly lower unlearning time. These results demonstrate that FINU provides an effective and efficient solution for full class unlearning in large-scale settings without requiring access to retain data.

C. Subclass Unlearning

Subclass unlearning is inherently challenging as compared to class level unlearning, as it requires forgetting specific classes without disrupting semantically similar ones within the same superclass. For example, in CIFARSuper20, one may aim to forget the class *baby* from the *people* superclass while retaining *boy*, *girl*, *man*, and *woman*.

We evaluate this fine-grained unlearning setting using ResNet-18 and ViT-B/16 on CIFARSuper20. Tables II and III summarize the results. Across both architectures, FINU achieves near-zero forget accuracy $\text{Acc}_{\mathcal{D}_f}$ while preserving high retain accuracy $\text{Acc}_{\mathcal{D}_r}$, indicating effective removal of the target subclass with minimal disruption to semantically related subclasses. In contrast, retain-data-free baselines such as Random Labeling and BDSH exhibit either incomplete forgetting or notable degradation in retain accuracy, while JiT remains competitive and performs slightly worse in ResNet18 as compared to the proposed approach. Compared to retain-data-based method SSD, FINU achieves comparable forgetting performance and better retain accuracy on ViT-B/16 with significantly lower unlearning time and consistently lower MIA scores, demonstrating strong privacy protection and efficiency in the subclass unlearning setting.

TABLE II: Sub-Class unlearning on CIFARSuper20 with finetuned ResNet18.

Method	ZS	$\text{Acc}_{\mathcal{D}_r}\uparrow$	$\text{Acc}_{\mathcal{D}_f}\downarrow$	MIA↓	Time↓
Original	X	85.41	89.0	0.952	–
Retrain	X	87.55	1.7	0.119	190
SSD [13]	X	81.53	0.0	0.069	18
Random Label [17]	✓	76.67	5.0	0.171	6
BDSH [7]	✓	75.38	3.0	0.186	13
JiT [8]	✓	70.58	1.0	0.156	8
FINU	✓	78.73	1.0	0.034	5

TABLE III: Sub-class unlearning on CIFARSuper20 with finetuned ViT-B/16.

Method	ZS	$\text{Acc}_{\mathcal{D}_r}\uparrow$	$\text{Acc}_{\mathcal{D}_f}\downarrow$	MIA↓	Time↓
Original	X	96.02	99.0	0.896	–
Retrain	X	96.00	8.0	0.126	1349
SSD [13]	X	92.69	0.0	0.129	123
Random Label [17]	✓	76.70	5.0	0.000	6
BDSH [7]	✓	96.01	97.0	0.214	30
JiT [8]	✓	95.52	0.0	0.000	12
FINU	✓	95.03	0.0	0.000	10

D. Sample Unlearning

In the sample-level unlearning setting, we randomly select 500 training samples to form the forget set $\mathcal{D}_f$, while the remaining samples constitute the retain set $\mathcal{D}_r$. The objective is to selectively remove the influence of these individual samples without degrading the model's generalization performance on the remaining data.

As shown in Table IV, sample-level unlearning is particularly challenging due to the sparse and distributed nature of the forget samples across classes. Compared to retain-data-free baselines, FINU achieves a stronger reduction in forget accuracy while maintaining competitive retain and test accuracy. While retraining provides a reference point for sample removal, it incurs higher computational cost. In contrast, FINU offers a favorable balance between selective forgetting, model utility, and privacy, demonstrating its effectiveness for fine-grained sample-level unlearning.

E. Assessment of FINU Unlearning via Grad-CAM

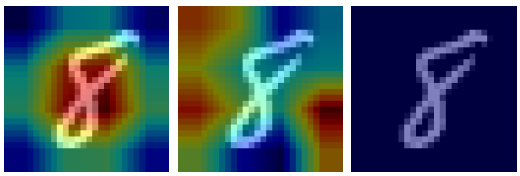
To qualitatively evaluate the effectiveness of the proposed FINU unlearning method, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [22] to visualize the regions of input images that most strongly influence model predictions. We analyze both the forget class (digit 8) and selected retain

TABLE IV: Random sample unlearning on CIFAR100 with finetuend ViT-B/16.

Method	**ZS**	**Acc**$_{\mathcal{D}_r}\uparrow$	**Acc**$_{\mathcal{D}_f}\downarrow$	**Acc**$_{\mathcal{D}_t}\uparrow$	**MIA**$\downarrow$
Original	X	99.98	100.0	92.26	0.92
Retrain	X	99.69	90.80	91.71	0.75
SSD	X	97.19	96.80	88.95	0.88
Random Label	✓	99.98	100.0	92.27	0.91
BDSH	✓	97.10	98.44	89.82	0.88
JiT	✓	97.98	98.27	**90.33**	0.89
FINU	✓	**98.10**	**94.60**	89.84	**0.85**

classes (digits 6 and 7) on the MNIST dataset using a ResNet18 architecture.

In the Grad-CAM visualizations shown in Fig. 1, blue regions indicate low activation and minimal contribution to the predicted class, whereas red regions denote high activation and strong influence on the classification decision. After FINU unlearning (Figs. 1c, the heatmaps for the forget class exhibit a dramatic shift: warm colors (red/yellow) are largely absent from the digit itself and are replaced by faint, low-intensity blue or purple patterns that form only a vague, ghost-like outline of the digit 8. The absence of high-importance regions on semantically meaningful parts of the digit indicates effective suppression of class-specific feature representations associated with the forgotten class. In contrast, for the retain classes, the Grad-CAM maps remain largely unchanged across the trained, retrained, and unlearned models. The dominant red activation regions continue to align with the digit strokes, demonstrating that FINU preserves discriminative features for retained classes while selectively removing information related to the forget class.



(a) Trained (b) Retrained (c) Unlearned

(d) Trained (e) Retrained (f) Unlearned

Fig. 1: GradCAM-based assessment of FINU unlearning on MNIST. The top row (a-c) corresponds to the forgotten class 8, while the bottom row (d-f) shows a retained class 6, comparing trained, retrained, and unlearned models.

F. Ablation Study

We conducted an ablation study to analyze the contribution of individual design choices in FINU and to quantify their impact on unlearning performance. All ablation experiments are performed on CIFAR-10 using a ResNet-18 architecture and under Class level unlearning setting.

1) Impact of Fisher-Guided Adaptive Parameter Masking: This ablation study evaluates the role of Fisher guidance and layer-wise adaptivity in FINU. We compare the proposed masking strategy against four alternatives: (i) no masking, where noise is applied uniformly to all parameters; (ii) random global masking, where a fixed fraction of parameters is randomly selected across the network; (iii) random layer-wise masking, where parameters are randomly selected independently within each layer; and (iv) Fisher-guided global masking, where parameters are selected based on Fisher information without layer-wise adaptivity.

TABLE V: Ablation on Fisher-guided adaptive parameter masking.

Masking Strategy	**Acc**$_{\mathcal{D}_f}\downarrow$	**Acc**$_{\mathcal{D}_r}\uparrow$	**Time**$\downarrow$
Trained Model	97.18	92.76	–
No Mask	0.00	12.02	10.21
Rand-Global	3.01	80.46	33.39
Rand-Layerwise	2.99	78.18	48.06
Fisher-Global	**0.00**	56.71	20.22
Fisher-Layerwise (FINU)	1.30	**90.11**	19.23

To ensure a fair comparison, we control for the total number of perturbed parameters across all strategies. Specifically, FINU is first applied with $(k_{\min}, k_{\max}) = (5\%, 15\%)$, resulting in 12.69% of parameters being selected. All global masking baselines perturb the same fraction of parameters, while the random layer-wise baseline follows the same $(k_{\min}, k_{\max})$ schedule without Fisher guidance.

Table V shows that applying noise without masking achieves complete forgetting but catastrophically degrades retain accuracy, confirming that unstructured perturbations are harmful. Random masking strategies, both global and layer-wise, improve retain performance but fail to achieve a strong forgetting–utility trade-off. Fisher-guided global masking achieves effective forgetting but causes substantial utility degradation, indicating that Fisher-based importance alone is insufficient. In contrast, FINU's Fisher-guided layer-wise masking achieves near-zero forget accuracy while preserving high retain accuracy and maintaining low unlearning time, demonstrating that both Fisher guidance and layer-wise adaptivity are essential for effective zero-shot unlearning.

2) Impact of Adaptive Top-k Range: FINU employs an adaptive layer-wise masking strategy in which the fraction of selected parameters increases with network depth. Specifically, for a network with L layers, the top-$k_\ell\%$ of parameters in layer ℓ are selected based on Fisher scores, where k_ℓ is linearly interpolated between a minimum and maximum value, denoted by $k_{\min}$ and $k_{\max}$, respectively. To study the effect of this design choice, we conduct an ablation by varying the $(k_{\min}, k_{\max})$ pair while keeping all other components of FINU fixed.

Table VI studies the effect of varying the adaptive top-k masking range in FINU. Small masking budgets lead to insuf-

TABLE VI: Ablation on adaptive top-k range.

$k_{\min}$	$k_{\max}$	$\text{Acc}_{\mathcal{D}_f}\downarrow$	$\text{Acc}_{\mathcal{D}_r}\uparrow$	Time↓
1%	5%	57.28	92.13	30.13
3%	10%	12.07	91.37	31.00
5%	**15%**	**1.19**	**90.16**	31.95
10%	25%	0.00	82.04	33.07
15%	30%	0.00	74.02	33.68

ficient forgetting, as too few parameters are perturbed. Increasing the masking range improves forgetting performance, with the default setting $(k_{\min}, k_{\max}) = (5\%, 15\%)$ achieving near-zero forget accuracy while preserving high retain accuracy. More aggressive masking results in perfect forgetting but substantially degrades model utility, indicating over-unlearning. Across all settings, the unlearning runtime remains largely stable, demonstrating that FINU is computationally robust to the choice of masking range.

3) Sensitivity analysis of hyperparameter λ: To assess the effect of the regularization strength λ in FINU, we vary λ while keeping all other settings fixed (Table VII). Increasing λ improves retain accuracy ($\text{Acc}_{\mathcal{D}r}$: 89.96 $\rightarrow$ 92.16) but degrades unlearning performance ($\text{Acc}\mathcal{D}_f$: 2.00 $\rightarrow$ 3.16), indicating a trade-off. Moderate values (0.01–1) achieve a favorable balance; in particular, $\lambda = 1$ yields strong retain performance (92.03) with limited degradation in forgetting (2.17).

TABLE VII: Ablation study on the sensitivity of λ.

λ	$\text{Acc}_{\mathcal{D}_r}\uparrow$	$\text{Acc}_{\mathcal{D}_f}\downarrow$
0	89.96	2.00
0.001	91.12	2.07
0.01	91.24	2.17
0.1	91.84	2.46
1	92.03	2.17
5	92.16	3.16

V. Conclusion & Outlook

In this paper, we have introduced FINU, a Fisher-guided Noise-based framework for zero-shot machine unlearning that operates without access to the retain dataset. FINU leverages the hierarchical structure of deep neural networks to selectively identify parameters most influential to the forget set using Fisher information and induces forgetting through structured, learnable noise injection, enabling effective removal of forget-set information while preserving generalizable knowledge. We evaluated FINU across class-level, subclass-level, and sample-level unlearning settings on CIFAR-100, CIFARSuper20, and ImageNet-1k using both convolutional and transformer-based architectures. Experimental results show that FINU achieves strong forgetting performance with competitive retain accuracy, improved resistance to membership inference attacks, and substantially lower unlearning time compared to existing zero-shot baselines. While FINU preserves model utility implicitly through structured masking and regularization rather than explicit retain-set optimization, extending the framework with tighter theoretical guarantees and to additional settings such as federated learning remains an important direction for future work.

References

[1] D. Arpit, S. Jastrzkebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 233–242.

[2] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," *ACM Trans. Intell. Syst. Technol.*, 2025.

[3] A. Mantelero, "The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'," *Comput. Law & Security Rev.*, vol. 29, no. 3, pp. 229–235, 2013.

[4] E. Goldman, "An introduction to the California Consumer Privacy Act (CCPA)," *Santa Clara Univ. Legal Studies Research Paper*, 2020.

[5] Office of the Privacy Commissioner of Canada, "Personal Information Protection and Electronic Documents Act," 2018.

[6] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *Proc. IEEE Symp. Security and Privacy (SP)*, 2021, pp. 141–159.

[7] M. Chen, W. Gao, G. Liu, K. Peng, and C. Wang, "Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7766–7775.

[8] J. Foster, K. Fogarty, S. Schoepf, Z. Dugue, C. Oztireli, and A. Brintrup, "An information theoretic approach to machine unlearning," *Trans. Mach. Learn. Res.*, 2025.

[9] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[11] R. Torkzadehmahani, R. Nasirigerdeh, G. Kaissis, D. Rueckert, G. K. Dziugaite, and E. Triantafillou, "Improved localized machine unlearning through the lens of memorization," *Trans. Mach. Learn. Res.*, 2025

[12] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Zero-shot machine unlearning," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2345–2354, 2023.

[13] J. Foster, S. Schoepf, and A. Brintrup, "Fast machine unlearning without retraining through selective synaptic dampening," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, no. 11, 2024, pp. 12043–12051.

[14] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Can bad teaching induce forgetting? Unlearning in deep networks using an incompetent teacher," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023.

[15] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, "Towards unbounded machine unlearning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[17] L. Graves, V. Nagisetty, V. Ganesh, "Amnesiac Machine Learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021.

[18] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., Univ. of Toronto, 2009.

[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 3320–3328.

[20] T. Eisenhofer, D. Riepel, V. Chandrasekaran, E. Ghosh, O. Ohrimenko and N. Papernot, "Verifiable and Provably Secure Machine Unlearning," 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), Copenhagen, Denmark, 2025, pp. 479-496, doi: 10.1109/SaTML64287.2025.00033.

[21] V. C. Gogineni and E. S. Nadimi, "Efficient knowledge deletion from trained models through layer-wise partial machine unlearning," Journal of Machine Learning Research, vol. 26, no. 245, pp. 1–33, 2025.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626.