CERTIFIED RETAIN DEGRADATION CONTROL FOR COUNTERFACTUAL GUIDED MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning can introduce new prediction errors on retained data, and searching over update configurations can amplify this risk through selection bias. Reliable deployment therefore requires a degradation guarantee that remains valid after configuration selection. We propose conformal machine unlearning configuration evaluation (CMUCE), a unified framework for certification and selection that combines deletion-efficacy screening with finite-sample control of positive prediction regret relative to a fixed reference model. For each realized deletion request, we establish a $1 - \delta$ guarantee that every accepted configuration has retained-population risk below a prescribed tolerance, even when candidates share calibration data and the final configuration is chosen from the accepted set. The construction combines Hoeffding–Bentkus bounds with family-wise testing and accommodates any calibration-independent candidate generator. We further characterize when counterfactual guidance improves the certificate: under explicit anchor-quality, gradient-alignment, and local score-transfer conditions, a paired one-step comparison yields a certificate-ordering bound whose calibration and anchor-yield error terms decay exponentially with their respective sample budgets. Under a Hellinger shift budget ρ, an additional risk margin of $\sqrt{2}\rho$ extends certification to the deployment distribution. Experiments on vision, text, and tabular benchmarks show that guided candidates can attain lower retain regret and tighter certificates, while joint selection identifies configurations meeting both deletion and retain-risk requirements.

1 INTRODUCTION

Machine unlearning (MU) updates a trained model after selected training records are removed (Bourtoule et al., 2021; Nguyen et al., 2025). A useful update must address the deletion request while preserving performance for the remaining population. This second objective becomes particularly important for targeted requests: unlearning can alter predictions and performance on the retained population (Chen et al., 2024; Oesterling et al., 2024). A deployment decision therefore needs evidence about both deletion efficacy and the additional errors introduced on retained data. Deletion-efficacy audits, including membership inference, and removal guarantees supplied by the base algorithm assess whether an update meets a specified deletion requirement (Tu et al., 2025; Koloskova et al., 2025). Retained accuracy describes aggregate utility, but it can hide a redistribution of errors: improved predictions on some examples can offset newly incorrect predictions on others. We instead measure the positive increase in per-example loss relative to a reference model. With the original model as reference, this score directly quantifies how often an update turns a previously correct retained prediction into an error.

Searching over update strengths, loss weights, counterfactual budgets, and anchor-quality thresholds can favor configurations whose observed retain scores underestimate their population risks. To support deployment under a user-specified tolerance, the guarantee must remain valid after this selection. It must also be paired with a deletion check: returning the original model gives zero additional error while leaving the request unaddressed. We introduce *conformal machine unlearning configuration evaluation* (CMUCE) to make this joint decision. Given candidates constructed before calibration, CMUCE combines a deletion audit with simultaneous tests of retained-population risk. The output is a set of configurations that pass the deletion criterion and have certified degradation below the chosen tolerance. Any subsequent choice within that set preserves the retain-side guarantee.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

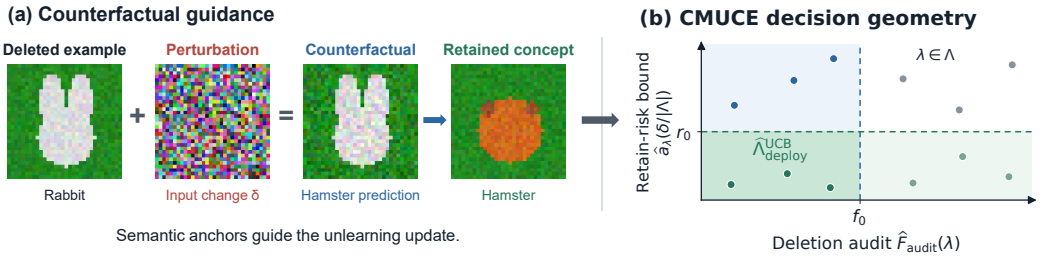

Figure 1: (a) Counterfactual anchors guide the unlearning update. (b) CMUCE compares deletion audits and simultaneous retain-risk upper bounds. Under Bonferroni calibration, the shaded $\widehat{\Lambda}^{\mathrm{UCB}}_{\mathrm{deploy}}$ gives a sufficient joint deployment rule: $\widehat{F}_{\mathrm{audit}}(\lambda) \le f_0$ and $\widehat{a}_\lambda(\delta/|\Lambda|) < r_0$.

This formulation accommodates counterfactual and other candidate generators within a common decision rule across unlearning methods.

Counterfactual guidance provides a concrete setting in which this rule is useful. Semantic anchors guide updates around deleted examples and can help preserve nearby retained predictions (Chen et al., 2024); Figure 1 illustrates the mechanism and the ensuing selection decision. We analyze when this guidance improves the numerical certificate. The comparison connects anchor yield, gradient alignment, and a local transfer from surrogate loss to the certified score. A separate shift extension translates a specified deployment-shift budget into a stricter calibration threshold.

Contributions. First, we formulate request-conditional, positive prediction regret as a deployment constraint paired with deletion efficacy. Second, we develop a modular certification-and-selection procedure using Hoeffding–Bentkus (HB) bounds and Learn-then-Test (LTT) (Angelopoulos et al., 2025); it provides both a fixed-candidate certificate and a family-wise guarantee that survives configuration selection. Third, we derive a local counterfactual comparison and a shift-sensitive acceptance rule, exposing how guidance quality and deployment assumptions affect the usable risk margin. Finally, experiments examine the joint feasible set, compare unlearning baselines, and study calibration, computational cost, and transfer across data modalities.

2 Related Work

Unlearning algorithms and evaluation. Machine unlearning spans efficient retraining schemes (Bourtoule et al., 2021), algorithmic removal guarantees for neural networks (Zhang et al., 2024; Koloskova et al., 2025), and methods evaluated through empirical forget–retain trade-offs (Nguyen et al., 2025; Tu et al., 2025). Representative methods include SCRUB (Kurmanji et al., 2023), UNSIR (Tarun et al., 2024), SalUn (Fan et al., 2024), MinNorm-OG (Block et al., 2025), SEMU (Sendera et al., 2025), and sparsity-aware unlearning (Jia et al., 2023). These approaches provide candidate updates with different computational and deletion properties. For configuration search, aggregate retain accuracy alone does not determine whether the additional errors introduced by the selected update remain below a specified population-risk tolerance.

Conformal inference and risk control. Conformal prediction provides marginal coverage under exchangeability (Angelopoulos & Bates, 2023); conformal risk control treats expected loss for suitable monotone families (Angelopoulos et al., 2024). LTT uses hypothesis testing to obtain high-probability risk control over candidate families (Angelopoulos et al., 2025). Family-wise procedures such as Bonferroni and Holm (Holm, 1979) account for testing multiple candidates even when their calibration scores are dependent. Applying these general tools to unlearning requires a bounded degradation score, a calibration population tied to the deletion request, and a decision rule that accommodates both configuration selection and deletion efficacy.

Counterfactual guidance and the effects of removal. Counterfactual examples have been used to identify and remove learned biases through machine unlearning (Chen et al., 2023). Counterfactual-guided unlearning further uses semantic alternatives to address data- and algorithm-level bias induced

by the forgetting process (Chen et al., 2024). Fair-unlearning research shows that removal also interacts with fairness objectives and develops guarantees that preserve fairness while unlearning data (Oesterling et al., 2024). These studies motivate examining which retained predictions change after an update and how the quality of counterfactual guidance affects that change. Connecting semantic anchor quality to a quantitative bound on newly introduced prediction errors requires an explicit account of the update dynamics and the evaluation score.

3 Problem Setup and Risk-Control Primitive

Let D_f be a realized deletion request and D_r the remaining training data. A pre-trained predictor h_θ and a finite, pre-specified grid Λ produce candidates $h_{\theta^u(\lambda)} = \mathsf{U}(h_\theta, D_f, D_r; \lambda)$. Let $\mathcal{F}$ contain D_f, all training and generator randomness, the candidate family, reference, score, deletion audit, thresholds, and testing rule, but no calibration observations. We require $Z_1, \ldots, Z_n \mid \mathcal{F} \overset{\text{i.i.d.}}{\sim} P_r(\cdot \mid D_f)$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. This conditional i.i.d. structure is the sampling assumption used by the HB guarantee below. Risks average over a fresh retained example for the realized candidate; retain-certification guarantees first hold conditionally on $\mathcal{F}$ and then, by averaging, on D_f.

Calibration protocol. Reserve calibration data independently from the start, excluding them from pretraining, unlearning, tuning, and counterfactual generation. Finalize the candidate grid, reference, and audit before inspecting these observations. For targeted deletion, sample from the retained population appropriate to the realized request. This protocol establishes the conditional sampling relationship above. An independent retained test split supports empirical diagnostics.

For a bounded score $s_\lambda : \mathcal{Z} \to [0,1]$, let $R(\lambda) = \mathbb{E}_{P_r}[s_\lambda(Z)]$ and $\widehat{R}(\lambda) = n^{-1}\sum_i s_\lambda(Z_i)$. For $\widehat{r} < a$, define the Hoeffding–Bentkus lower-tail function (Angelopoulos et al., 2025)

$$B_n(\widehat{r}, a) = \min\Big\{e^{-nh(\widehat{r},a)},\ e\ \Pr[\mathrm{Bin}(n,a) \le \lceil n\widehat{r}\rceil]\Big\}, \quad h(u,v) = u\log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v}, \tag{1}$$

Set $B_n(\widehat{r}, a) = 1$ for $\widehat{r} \ge a$, with $n \ge 1$ and $\delta \in (0,1)$. Interpret binary KL by its continuous extended-value convention, including $h(0,a) = -\log(1-a)$ and $h(u,1) = +\infty$ for $u < 1$. Inversion gives a fixed-candidate upper bound. Testing $R(\lambda) \ge r_0$, $r_0 \in (0,1)$, with family-wise error control gives a post-selection valid set. These two uses answer complementary questions: inversion quantifies risk for a candidate specified before calibration, and simultaneous testing supports selection from a candidate family. Simultaneous testing preserves the retain-side guarantee for any downstream choice from the returned set.

4 Joint Deletion-Efficacy and Retain-Degradation Control

4.1 Two complementary constraints

Fix a reference θ^{ref} independently of D_{cal}. The default pre-unlearning reference answers: what additional retained error did this request induce? For $z = (x, y)$, define

$$s_{\mathrm{ret}}(\lambda; z) = \big[\ell_{01}(\theta^u(\lambda); z) - \ell_{01}(\theta^{\mathrm{ref}}; z)\big]_+ \in [0,1], \qquad R_{\mathrm{ret}}(\lambda) = \mathbb{E}_{P_r}[s_{\mathrm{ret}}(\lambda; Z)]. \tag{2}$$

For 0–1 loss, $s_{\mathrm{ret}} = 1$ precisely when the reference predicts correctly and the updated model predicts incorrectly. Improvements on other examples do not offset these newly introduced errors. The score therefore complements retained accuracy with an explicit measure of the population affected adversely by the update. A retain-only retrained reference instead compares the candidate with a retraining target; the reference is fixed before calibration. With $\theta^{\mathrm{ref}} = \theta$, no-op attains $R_{\mathrm{ret}} = 0$.

Let $\widehat{F}_{\mathrm{audit}}(\lambda)$ be a separately specified deletion-efficacy statistic computed without D_{cal}, and set

$$\Lambda_{\mathrm{eff}} = \{\lambda \in \Lambda : \widehat{F}_{\mathrm{audit}}(\lambda) \le f_0\}. \tag{3}$$

For class deletion we use deleted-class accuracy; for sample deletion we use a pre-specified membership-inference advantage. These empirical audits specify the operational deletion requirement (Tu et al., 2025); a removal certificate from the base unlearner can also serve as the criterion.

Algorithm 1 CMUCE certification-and-selection layer

1: **Input:** pre-specified Λ and FWER rule $\mathcal{A}$, candidates $\{\theta^u(\lambda)\}$, D_{cal}, deletion audits, r_0, f_0, δ, fixed θ^{ref}.
2: Form Λ_{eff} by Eq. (3) without using D_{cal}.
3: **for** $\lambda_j \in \Lambda$ **do**
4: Evaluate Eq. (2) on D_{cal} and compute Eq. (5).
5: **end for**
6: Apply $\mathcal{A}$ to all $|\Lambda|$ p-values at level δ; **return** $\Lambda_{\mathrm{eff}} \cap \widehat{\Lambda}_{r_0}$ and labeled fixed-candidate bounds.

4.2 Certification and finite-grid selection

Compute $\widehat{R}_{\mathrm{ret}}(\lambda) = N_{\mathrm{cal}}^{-1} \sum_i s_{\mathrm{ret}}(\lambda; Z_i)$. For fixed λ, define

$$\widehat{a}_\lambda(\delta) = \inf\{a \in [\widehat{R}_{\mathrm{ret}}(\lambda), 1] : B_{N_{\mathrm{cal}}}(\widehat{R}_{\mathrm{ret}}(\lambda), a) \le \delta\}, \tag{4}$$

with value 1 if the set is empty.

Proposition 4.1 (Fixed-candidate retain certificate)**.** *If λ is fixed independently of D_{cal}, then, conditional on D_f,* $\Pr[R_{\mathrm{ret}}(\lambda) \le \widehat{a}_\lambda(\delta)] \ge 1 - \delta$.

When the same calibration set is used to search $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$, test $H_j : R_{\mathrm{ret}}(\lambda_j) \ge r_0$ with

$$p_j = \begin{cases} B_{N_{\mathrm{cal}}}(\widehat{R}_{\mathrm{ret}}(\lambda_j), r_0), & \widehat{R}_{\mathrm{ret}}(\lambda_j) < r_0, \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Apply Bonferroni, Holm (Holm, 1979), or another pre-specified valid FWER procedure $\mathcal{A}$ at level δ and let $\widehat{\Lambda}_{r_0} = \{\lambda_j : j \in \mathcal{A}(p_1, \ldots, p_N)\}$. These procedures control the probability of rejecting any true unsafe null, including when candidate scores are dependent.

Proposition 4.2 (Post-selection retain control)**.** *Conditional on D_f,* $\Pr[\forall \lambda \in \widehat{\Lambda}_{r_0},\ R_{\mathrm{ret}}(\lambda) < r_0] \ge 1 - \delta$.

Why the intersection remains valid. On the FWER event, every member of $\widehat{\Lambda}_{r_0}$ is retain-safe simultaneously. Intersecting with any deletion-effective set formed without the calibration sample only removes candidates, so it cannot introduce a retain-unsafe member. No independence among candidate scores or p-values is required by Bonferroni or Holm. The resulting set pairs the chosen deletion criterion with simultaneous retain-risk control. CMUCE returns $\widehat{\Lambda}_{\mathrm{deploy}} = \Lambda_{\mathrm{eff}} \cap \widehat{\Lambda}_{r_0}$. Because this is a subset of $\widehat{\Lambda}_{r_0}$, Proposition 4.2 holds for every returned candidate without another multiplicity penalty. An empty set prompts abstention. Algorithm 1 summarizes the complete certification-and-selection procedure; Appendix B proves both retain-risk guarantees and the simultaneous numerical-bound construction.

Using the output. Any selection from $\widehat{\Lambda}_{\mathrm{deploy}}$ inherits $R_{\mathrm{ret}} < r_0$ on the simultaneous event. A practitioner can therefore choose a faster candidate or favor auxiliary utility within the accepted set. For numerical bounds that remain valid after selection, Bonferroni inversion gives $\{\widehat{a}_\lambda(\delta/N) : \lambda \in \Lambda\}$ with simultaneous coverage at least $1 - \delta$. Nominal bounds $\widehat{a}_\lambda(\delta)$ characterize candidates specified before calibration. These outputs specify which candidate can be selected and which risk bound can be reported.

Planning calibration. For a planned empirical risk ceiling $r_{\mathrm{plan}} < r_0$, the conservative Hoeffding–Bonferroni sufficient condition is

$$N_{\mathrm{cal}} \ge \frac{\log(N/\delta)}{2(r_0 - r_{\mathrm{plan}})^2}. \tag{6}$$

Any realized score average at most r_{plan} then passes the Bonferroni threshold. The required sample size grows logarithmically with candidate count and quadratically with the inverse tolerance margin. HB can be sharper near zero. When no calibration example incurs regret, its Hoeffding term is $(1 - r_0)^{N_{\mathrm{cal}}}$, so $N_{\mathrm{cal}} \ge \lceil \log(\delta/N) / \log(1 - r_0) \rceil$ suffices. A new search prompted by calibration feedback calls for fresh calibration data or a sequentially valid procedure.

5 When Counterfactual Guidance Improves Certification

The preceding procedure certifies any candidate generator satisfying the setup. We now connect counterfactual guidance to a smaller numerical certificate through a local comparison of two updates. The analysis links valid-anchor yield, retained-loss improvement, and transfer to the bounded certification score. Appendix C details the generator assumptions, alignment, and proofs.

Let $L_r(\theta) = \mathbb{E}_{P_r}[\ell_{\mathrm{sur}}(\theta; Z)]$ be L-smooth and $g_r(\theta) = \nabla L_r(\theta)$. For $\eta > 0$ and guidance weight $w_{\mathrm{cf}} > 0$, compare

$$\theta^{\mathrm{std}} = \theta - \eta g_f, \qquad \theta^{\mathrm{cf}} = \theta^{\mathrm{std}} - \eta w_{\mathrm{cf}} g_{\mathrm{cf}}. \tag{7}$$

Assume $\|g_f\| \le B_f$ and $\|g_{\mathrm{cf}}\| \le B_{\mathrm{cf}}$. Here g_f is the base update direction, and both directions are evaluated at the initial θ. Appendix A specifies the candidate objective and its gradient decomposition. For each fixed input x and generator, independent seeds produce $Q_x \in [0, 1]$, with zero assigned to infeasible or failed outputs. Assume $\mathbb{E}[Q_x \mid x] \ge \mu_q > \tau_q > 0$ and $\mathrm{Var}(Q_x \mid x) \le V_{\mathrm{gen}}^2$. The constants hold uniformly over the inputs considered.

More explicitly, Cantelli yields

$$\begin{aligned} p_{\mathrm{val}}(x) := \Pr(Q_x \ge \tau_q \mid x) \ge \underline{p} := 1 - \frac{V_{\mathrm{gen}}^2}{V_{\mathrm{gen}}^2 + (\mu_q - \tau_q)^2}, \\ \Pr\left(\frac{N_{\mathrm{valid}}}{N_{\mathrm{cf}}} < \nu \,\middle|\, x \right) \le e^{-2N_{\mathrm{cf}}(\underline{p} - \nu)^2}. \end{aligned} \tag{8}$$

for $N_{\mathrm{cf}} \ge 1$ and $0 < \nu < \underline{p}$. A larger quality margin and a lower generator variance support a larger guaranteed valid-anchor fraction. Appendix C.2 gives the conditional proof.

Let $\mathcal{E}_{\mathrm{align}} \in \mathcal{F}$ have probability at least π_{align} and let $\mathcal{E}_{\mathrm{val}} = \{N_{\mathrm{valid}}/N_{\mathrm{cf}} \ge \nu\} \in \mathcal{F}$. On their intersection, assume directly

$$\langle g_r(\theta), g_{\mathrm{cf}} \rangle \ge d^+ \kappa \nu, \qquad d^+, \kappa, \nu > 0, \tag{9}$$

One sufficient factorization is cosine alignment at least d^+ together with norm product at least $\kappa\nu$. This accounts for both the direction and magnitude of the guidance signal. Smoothness at θ^{std} yields

$$L_r(\theta^{\mathrm{std}}) - L_r(\theta^{\mathrm{cf}}) \ge \eta w_{\mathrm{cf}} \langle g_r(\theta), g_{\mathrm{cf}} \rangle - L\eta^2 w_{\mathrm{cf}} B_f B_{\mathrm{cf}} - \tfrac{L\eta^2 w_{\mathrm{cf}}^2}{2} B_{\mathrm{cf}}^2. \tag{10}$$

Define the resulting margin

$$M_\nu = \eta w_{\mathrm{cf}} d^+ \kappa \nu - L\eta^2 w_{\mathrm{cf}} B_f B_{\mathrm{cf}} - \tfrac{L\eta^2 w_{\mathrm{cf}}^2}{2} B_{\mathrm{cf}}^2. \tag{11}$$

Its positive term captures the aligned signal; the quadratic terms account for movement under the base update and curvature along the guidance direction. In particular, $M_\nu > 0$ when

$$0 < \eta < \frac{d^+ \kappa \nu}{L B_f B_{\mathrm{cf}} + (L w_{\mathrm{cf}}/2) B_{\mathrm{cf}}^2}, \tag{12}$$

for a positive denominator. This identifies local step sizes for which aligned guidance dominates curvature.

For bounded scores $s_{\mathrm{std}}, s_{\mathrm{cf}} \in [0, 1]$, the required bridge is the local score-transfer condition

$$\mathbb{E}_{P_r}[s_{\mathrm{std}}(Z) - s_{\mathrm{cf}}(Z)] \ge c_s M_\nu =: m_\nu > 0 \quad \text{on } \mathcal{E}_{\mathrm{align}} \cap \mathcal{E}_{\mathrm{val}}. \tag{13}$$

Theorem 5.1 (Conditional certificate improvement)**.** *Under the conditional i.i.d. setup and Eqs.* (7)–(13)*, with fixed constants $M_\nu > 0$ and $c_s > 0$, certificates at a common confidence level on the same calibration sample satisfy*

$$\Pr[\widehat{a}_{\mathrm{cf}}(\delta) \le \widehat{a}_{\mathrm{std}}(\delta)] \ge 1 - e^{-N_{\mathrm{cal}} m_\nu^2/2} - (1 - \pi_{\mathrm{align}}) - \Pr(N_{\mathrm{valid}}/N_{\mathrm{cf}} < \nu). \tag{14}$$

Each fixed-candidate bound remains marginally valid by Proposition 4.1; the pair is jointly valid with probability at least $1 - 2\delta$.

Interpretation and proof. The three terms quantify calibration fluctuation, the alignment event, and valid-anchor yield. Given budgets $\varepsilon_c, \varepsilon_a, \varepsilon_v \in (0,1)$ with $\pi_{\text{align}} \geq 1-\varepsilon_a$, choosing $N_{\text{cal}} \geq 2m_\nu^{-2}\log(1/\varepsilon_c)$ and $N_{\text{cf}} \geq [2(\underline{p}-\nu)^2]^{-1}\log(1/\varepsilon_v)$ gives ordering probability at least $1-\varepsilon_c-\varepsilon_a-\varepsilon_v$, under the same local conditions. The proof conditions on all candidate information, applies Hoeffding to paired score differences, and uses monotonicity of HB inversion (Appendix C.4). The result explains a one-step mechanism by which useful guidance can improve certification; the risk-control propositions apply throughout the candidate family.

A concrete score-transfer regime. For a pre-fixed $M > 0$, define the bounded surrogate score $s_{\text{sm}}(\theta, z) = \min\{1, [\ell_{\text{sur}}(\theta; z) - \ell_{\text{sur}}(\theta^{\text{ref}}; z)]_+/M\}$. If both candidates' untruncated regrets lie in $(0, M)$ almost surely, clipping is inactive and the expected score gap equals the surrogate-risk gap divided by M, giving $c_s = 1/M$. For the default 0–1 score, Eq. (13) is a separate assumption about changes in classification errors. This connects the optimized loss to the certified score.

6 Shift-Aware Sensitivity Certificate

Let D be the calibration law, Q a deployment law, and suppose a fixed external budget justifies $H(D,Q) \leq \rho$, where $H^2(D,Q) = \frac{1}{2}\int(\sqrt{dD/d\mu} - \sqrt{dQ/d\mu})^2\, d\mu$ for a dominating measure μ. For every $f : \mathcal{Z} \to [0,1]$, $|\mathbb{E}_Q f - \mathbb{E}_D f| \leq \text{TV}(D,Q) \leq \sqrt{2}\rho$.

Theorem 6.1 (Drift-aware retain certificate)**.** *For a fixed λ chosen independently of D_{cal}, let $\widehat{a}_D^{\text{HB}}(\delta)$ be the valid Hoeffding–Bentkus upper bound under D. If $H(D,Q) \leq \rho$, then with probability at least $1-\delta$,*

$$\mathbb{E}_Q[s_{\text{ret}}(\lambda; Z)] \leq \widehat{a}_\lambda^{\text{shift}}(\rho, \delta) := \min\{1, \widehat{a}_D^{\text{HB}}(\delta) + \sqrt{2}\rho\}. \tag{15}$$

For a deployment target r_0, let $b_\rho = r_0 - \sqrt{2}\rho$. If $b_\rho \leq 0$, return the empty set. Otherwise test the source nulls $R_D(\lambda) \geq b_\rho$ using Eq. (5) with r_0 replaced by b_ρ and control FWER at δ. All accepted candidates then satisfy $R_Q(\lambda) < r_0$ simultaneously with probability at least $1-\delta$. Simultaneous numerical bounds give the alternative sufficient rule $\widehat{a}_\lambda(\delta/N) + \sqrt{2}\rho < r_0$. This extension expresses distribution shift as a risk budget. A valid source certificate $a_D < r_0$ leaves headroom $\rho_{\max} = (r_0 - a_D)/\sqrt{2}$, with strict acceptance for $\rho < \rho_{\max}$. A selected candidate uses a simultaneous source bound. Reporting this boundary makes the assumed shift tolerance explicit; estimating the actual radius requires external evidence. When labeled deployment observations become available, they can instead support direct recertification under the new population. Appendix D proves the shift certificate and derives the anchor-yield bound and empirical-Bernstein comparison.

7 Evaluation

The main experiments examine counterfactual effects on certificate values, joint deletion/retain-risk selection, sensitivity to calibration budget and reference choice, transfer across data modalities, and certification cost. Appendix E adds protocol details, noisy-anchor tests, alternative vision backbones, tabular configuration sweeps, runtime components, and HB-certified non-counterfactual baselines.

7.1 Protocol

We evaluate CIFAR-100 (Krizhevsky, 2009) and CUB-200 (Wah et al., 2011) with ResNet-18 (He et al., 2016) for class- and sample-deletion requests. The retained data pool is split 80:10:10 into training, calibration, and test subsets. CF-guided updates use gradient descent (learning rate 10^{-4}, batch size 128, seed 42). Calibration is reserved before pretraining and excluded from unlearning, counterfactual generation, tuning, and deletion auditing. The reference, score, candidate family, and testing rule are fixed before calibration inspection. Appendix A specifies the counterfactual-guided candidate objective. The grid varies $N_{\text{cf}} \in \{0, 10, 50, 100\}$ and $\tau_q, \alpha, \beta, \gamma$ are 0.2, giving 52 configurations (Appendix E). Unless stated otherwise, $\delta = 0.1$ and $\theta^{\text{ref}} = \theta$. For each dataset and deletion setting, Bonferroni tests the entire 52-candidate family:

$$\widehat{\Lambda}_{r_0} = \{\lambda_j \in \Lambda : p_j \leq 0.1/52\}, \qquad \widehat{\Lambda}_{\text{deploy}} = \widehat{\Lambda}_{r_0} \cap \Lambda_{\text{eff}}. \tag{16}$$

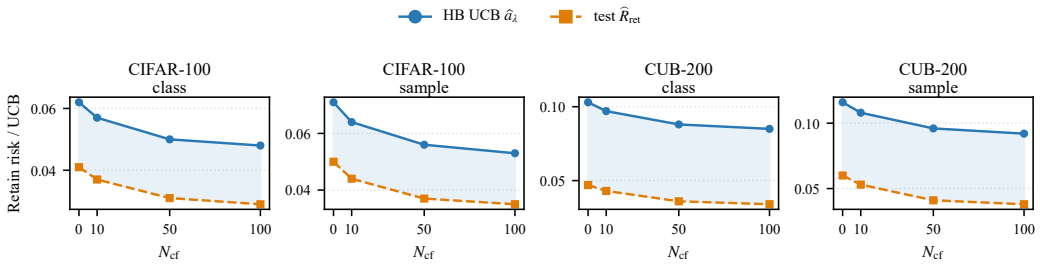

Figure 2: Fixed-candidate retain regret and HB upper bounds as the counterfactual budget increases. The shaded region is the difference between the UCB and independent-test regret.

Table 1: Twenty-class deletion: RA/FA are percentages; R/UCB are retain-regret probabilities.

Method	CIFAR RA/FA	CIFAR R/UCB	CUB RA/FA	CUB R/UCB
Retrain	78.2/0.0	.008/.028	82.3/0.0	.017/.072
FT-Unlearn	75.6/7.1	.034/.054	79.1/9.0	.049/.104
SCRUB (Kurmanji et al., 2023)	74.9/6.3	.041/.061	78.6/8.2	.054/.109
UNSIR (Tarun et al., 2024)	75.4/6.4	.038/.057	78.8/8.3	.052/.109
SalUn (Fan et al., 2024)	76.2/5.7	.029/.045	79.7/7.2	.045/.103
MinNorm-OG (Block et al., 2025)	76.4/5.9	.027/.049	79.1/6.9	.044/.099
SEMU (Sendera et al., 2025)	76.7/5.1	.031/.054	79.3/7.2	.047/.105
ℓ_1-Sparse (Jia et al., 2023)	76.3/5.2	.027/.048	79.4/6.7	.043/**.096**
CF-guided + CMUCE	**76.8**/5.6	**.022/.042**	**79.9**/7.3	**.041/.096**

The family contains all 52 configurations before deletion filtering, with a separate family-wise guarantee for each setting at its fixed tolerance. Class deletion uses deleted-class accuracy (FA) $\leq 10\%$; sample deletion uses membership-inference advantage ≤ 0.05. We report retained accuracy (RA), independent-test retain regret (R), and fixed-candidate HB upper confidence bounds (UCBs) at $\delta = 0.1$. Joint-set selection uses Eq. (16), accounting for the full configuration search.

7.2 Counterfactual Guidance and Certificate Values

Figure 2 compares the HB bound with the same bounded score on an independent test split as counterfactual coverage varies. The UCB stays above the test proxy in every displayed setting and tightens with N_{cf}. For example, CIFAR class deletion changes from risk/UCB $.041/.062$ at $N_{\text{cf}} = 0$ to $.029/.048$ at $N_{\text{cf}} = 100$; CUB sample deletion changes from $.060/.116$ to $.038/.092$. Guidance yields lower observed degradation together with tighter numerical certificates in these settings.

7.3 Joint Selection and Comparison with Unlearning Baselines

Table 1 reports matched 20-class deletion, with $N_{\text{cf}} = 50$ for the CF-guided candidate. RA is retained accuracy, FA is deleted-class accuracy (lower is stronger deletion), R is independent-test retain regret, and UCB is the Hoeffding–Bentkus certificate. All approximate methods use matched gradient-step counts and optimizer settings; wall-clock times, including counterfactual generation, are reported in Appendix E.6. Figure 3 shows how the separate deletion and retain-risk criteria determine the deployable set. At $r_0 = 0.05$, the reported triplets $(|\Lambda_{\text{eff}}|, |\widehat{\Lambda}_{r_0}|, |\widehat{\Lambda}_{\text{deploy}}|)$ are $(31, 23, 18)$, $(28, 19, 15)$, $(25, 16, 12)$, and $(23, 14, 10)$ for CIFAR-class, CIFAR-sample, CUB-class, and CUB-sample, respectively. Class deletion uses FA≤ 10; sample deletion uses MIA advantage ≤ 0.05. Each setting offers multiple jointly admissible candidates for selection by computational cost or auxiliary utility. Among approximate methods, CF-guided + CMUCE has the lowest retain risk on both datasets and the tightest CIFAR UCB; on CUB its UCB ties ℓ_1-Sparse. Its FA remains within the deletion threshold. This combination preserves strong retained utility while meeting the deployment deletion criterion; retraining provides a reference for the attainable forget–retain trade-off.

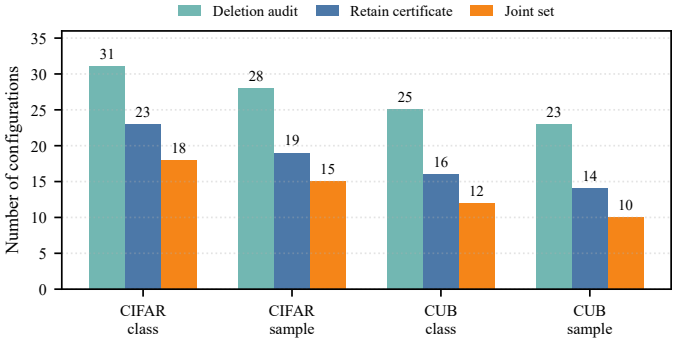

Figure 3: Deletion-effective, retain-certified, and joint configuration sets at $r_0 = 0.05$. Retain certification uses Bonferroni over the full grid in Eq. (16).

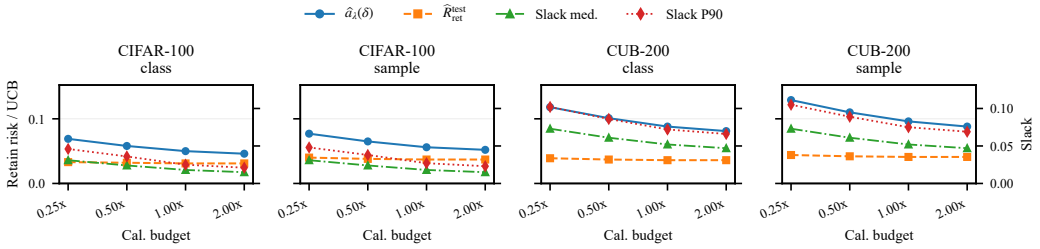

Figure 4: Calibration-budget sensitivity: HB upper bounds and UCB–test-risk slack decrease with calibration size, while retain-test risk changes little. Slack med./P90 denote its median/90th percentile.

7.4 Calibration and Reference Diagnostics

Figure 4 brings the calibration-budget comparison alongside the planning bound in Eq. (6). Across budgets from $0.25\times$ to $2.0\times$ the default size, the reported retain-test risks vary little, while the HB upper bounds and their slack decrease. More calibration data therefore reduce the uncertainty around a candidate's risk. Appendix E.1 describes the comparison protocol. The zero-regret calculation gives a useful endpoint: for $N = 52$, $\delta = 0.1$, and $r_0 = 0.05$, 122 zero-regret calibration observations suffice for Bonferroni acceptance using the Hoeffding term. This is a planning example; realized nonzero scores are evaluated by the full HB rule. Reference choice changes the comparison being made. Replacing the original reference with a retain-only retrained model changes risk/UCB from $.022/.042$ to $.027/.048$ on CIFAR and from $.041/.096$ to $.047/.104$ on CUB. The former asks how much degradation the update introduces relative to the deployed model; the latter asks how closely its predictions preserve the successes of retraining. Reporting the reference alongside the certificate gives these values a clear interpretation. Paired surrogate/score diagnostics provide a descriptive view of the local relationship in Section 5. Across 156 standard/CF pairs per dataset, the same-sign rates are 79% on CIFAR and 76% on CUB, with regression slopes 0.18 and 0.15. These associations motivate the score-transfer condition while leaving its uniform form explicit in Theorem 5.1. The complementary robustness and backbone comparisons are organized in Appendices E.2 and E.3; Appendix E.7 compares the same HB evaluation applied to non-counterfactual unlearners.

7.5 Transfer Across Data Modalities

Figure 5 extends the evaluation to text and tabular data. On AG News (Zhang et al., 2015), DistilBERT (Sanh et al., 2019) is the base predictor and FLAN-T5-small (Chung et al., 2024) generates counterfactual rewrites. We evaluate class deletion and uniform sample deletion. On Adult (Becker & Kohavi, 1996) and German (Hofmann, 1994), a three-layer DNN uses C-CHVAE anchors (Pawelczyk et al., 2020) for a group-targeted request: 30% of training records in the female group are removed. This retains the feature schema while changing the retained population's composition. Appendix E.4

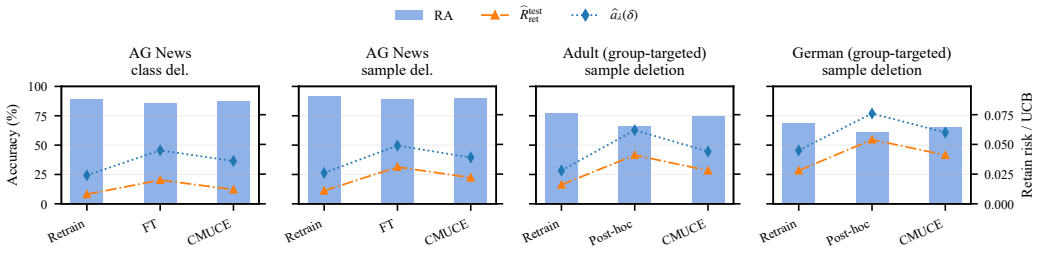

Figure 5: Text and tabular pilots. CMUCE denotes a CF-guided candidate with retain-risk certification; FT/Post-hoc denote the unlearning baselines. Each setting reports retained accuracy, independent-test regret, and a fixed-candidate HB upper bound.

Table 2: Seconds per configuration; update-step budgets are matched. Parentheses include HB.

Method	CIFAR class	CIFAR sample	CUB class	CUB sample
FT-Unlearn	35.8 (38.5)	33.2 (35.8)	49.5 (53.3)	46.2 (49.8)
SCRUB	47.1 (49.8)	44.3 (46.9)	64.5 (68.3)	60.8 (64.4)
SalUn	44.6 (47.3)	41.7 (44.3)	59.3 (63.1)	56.3 (59.9)
MinNorm-OG	40.6 (43.3)	37.1 (39.7)	57.3 (61.1)	53.9 (57.5)
SEMU	42.8 (45.5)	39.2 (41.8)	59.9 (63.7)	54.1 (57.7)
ℓ_1-Sparse	36.9 (39.6)	34.8 (37.4)	51.6 (55.4)	46.9 (50.5)
CF-guided	50.5 (53.2)	47.9 (50.5)	68.3 (72.1)	65.1 (68.7)

details the deletion regimes, anchor generation, and modality-specific quality filtering. Across these settings, the CF-guided candidates preserve higher retained accuracy than the post-hoc baselines, with lower observed regret and tighter UCBs. The certification input remains the same bounded per-example score even as the predictor and anchor generator change. This gives a common interpretation to the results: the UCB quantifies additional retained error relative to the fixed reference in each task. Appendix E.5 reports tabular configuration sweeps and descriptive screening counts showing how candidate quality varies with generation budget and loss weights.

7.6 Computational cost

Table 2 reports per-configuration runtimes before and after adding HB certification, with matched update-step budgets and optimizer settings. For example, the CF-guided CIFAR-class entry increases from 50.5 to 53.2 seconds: 42.1 seconds are spent on the base update, 8.4 on counterfactual generation, and 2.7 on calibration. Figure 9 in Appendix E.6 shows this three-part breakdown and peak GPU memory for all four vision settings. Scoring all $|\Lambda|$ candidates on N_{cal} retained examples requires $|\Lambda|N_{\text{cal}}$ candidate–example evaluations, followed by HB inversion and family-wise testing of the resulting scalar risk estimates and p-values. Serial grid evaluation sums the per-configuration costs; candidate evaluations can also run in parallel.

8 Conclusion and Future Directions

CMUCE turns retain-side degradation into an explicit, selection-valid deployment constraint for machine unlearning. Combining this constraint with a deletion audit yields a joint set from which practitioners can choose according to utility or cost. The local analysis connects useful counterfactual guidance to tighter certificates, while the shift extension expresses deployment uncertainty through a transparent risk budget. Experiments illustrate the resulting decisions across unlearning methods and data modalities. Future work can integrate algorithmic removal certificates into the joint rule, develop sequential calibration for repeated requests, and extend the local comparison to multi-step updates and classification-margin conditions. Richer group-specific scores and data-driven shift assessment would broaden the framework's deployment scope.

AI USE STATEMENT

Generative AI assisted writing and translation. The authors are responsible for the final content.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://archive.ics.uci.edu/dataset/2/adult.

Jacob Block, Aryan Mokhtari, and Sanjay Shakkottai. Machine unlearning under overparameterization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 109307–109348, 2025. doi: 10.52202/085713-3646. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9d7fe07d48d198449d0dbefe86263fdb-Abstract-Conference.html.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model DeBias with machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14516–14539, 2023. doi: 10.52202/075280-0639. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2ecc80084c96cc25b11b0ab995c25f47-Abstract-Conference.html.

Ziheng Chen, Jia Wang, Jun Zhuang, Abbavaram Gowtham Reddy, Fabrizio Silvestri, Jin Huang, Kaushiki Nag, Kun Kuang, Xin Ning, and Gabriele Tolomei. Debiasing machine unlearning with counterfactual examples. *arXiv preprint arXiv:2404.15760*, 2024. URL https://arxiv.org/abs/2404.15760.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ec4d2e436794d1bf55ca83f5ebb31887-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Hans Hofmann. Statlog (German credit data). UCI Machine Learning Repository, 1994. URL https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51584–51605, 2023. doi: 10.52202/075280-2246. URL https://papers.neurips.cc/paper_files/paper/2023/hash/a204aa68ab4e970e1ceccfb5b5cdc5e4-Abstract-Conference.html.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31275–31298, 2025. URL https://proceedings.mlr.press/v267/koloskova25a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1957–1987, 2023. doi: 10.52202/075280-0095. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract-Conference.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025. doi: 10.1145/3749987. URL https://doi.org/10.1145/3749987.

Alex Oesterling, Jiaqi Ma, Flavio Calmon, and Himabindu Lakkaraju. Fair machine unlearning: Data removal while mitigating disparities. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3736–3744. PMLR, 2024. URL https://proceedings.mlr.press/v238/oesterling24a.html.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132. Association for Computing Machinery, 2020. doi: 10.1145/3366423.3380087. URL https://doi.org/10.1145/3366423.3380087.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Damian Rymarczyk. SEMU: Singular value decomposition for efficient machine unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53843–53866, 2025. URL https://proceedings.mlr.press/v267/sendera25a.html.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2024. doi: 10.1109/TNNLS.2023.3266233.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Yiwen Tu, Pingbang Hu, and Jiaqi Ma. A reliable cryptographic framework for empirical machine unlearning evaluation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131705–131736, 2025. doi: 10.52202/085713-4384. URL https://papers.nips.cc/paper_files/paper/2025/hash/beb290ec14261c16b9269a256b55f029-Abstract-Conference.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58800–58818, 2024. URL https://proceedings.mlr.press/v235/zhang24l.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

A Additional Details of the CMUCE Protocol

This appendix provides the candidate-generation protocol (Appendix A), proofs of retain-risk certification and configuration selection (Appendix B), the analysis of counterfactual guidance (Appendix C), and extensions under distribution shift (Appendix D). Appendix E provides supplementary calibration and implementation details, together with robustness, transfer, and efficiency results. Throughout this appendix, $h_{\boldsymbol{\omega}}$ denotes the logit-valued network, $f_{\boldsymbol{\omega}}(x) = \mathrm{softmax}(h_{\boldsymbol{\omega}}(x))$ denotes the predictive distribution, and $\hat{y}_{\boldsymbol{\omega}}(x) = \arg\max_y f_{\boldsymbol{\omega}}(y \mid x)$ denotes the predicted label.

A.1 Counterfactual-guided base-update protocol

CMUCE accepts candidates from any generator that satisfies the calibration-separation conditions. We use a predictive-alignment variant inspired by Chen et al. (2024). A frozen teacher supplies counterfactual targets, and a student learns to match those targets on the original forget inputs. This formulation uses ordinary predictive distributions. Let $h_{\boldsymbol{\omega}}$ denote the teacher and $h_{\boldsymbol{\omega}^u}$ the student, whose encoder and prediction head are both updated. Candidate generation uses the training splits of $\mathcal{D}_f$ and $\mathcal{D}_r$.

Counterfactual anchors. For each forget instance $X_j \in \mathcal{D}_f$, a counterfactual anchor is a nearby input that preserves task-relevant semantics while changing the prediction of the pre-unlearning model. Let $(\mathcal{X}, d)$ be the input space with distance d, and let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a bounded semantic-consistency surrogate. A counterfactual anchor X_j^{cf} can be written as a feasible solution of

$$X_j^{\mathrm{cf}} \in \arg\min_{x' \in \mathcal{X}} d(X_j, x') \quad \text{s.t.} \quad \arg\max h_{\boldsymbol{\omega}}(X_j) \neq \arg\max h_{\boldsymbol{\omega}}(x'), \quad c(X_j, x') \leq \varepsilon_{\mathrm{sem}}. \tag{A.1}$$

A fixed generator approximates this constrained search and returns the feasible anchors it finds; the returned set may be empty. The generator, feasibility checks, distance, semantic surrogate, and stopping rule are fixed using training data before calibration.

Quality scoring and filtering. Given a forget minibatch, the generator produces a candidate set

$$Z_{\mathrm{cf}}(X_j) = \{X_{j,1}^{\mathrm{cf}}, \ldots, X_{j,N_{\mathrm{cf}}}^{\mathrm{cf}}\}, \tag{A.2}$$

Here N_{cf} counts all generation attempts, including failed and infeasible attempts. The collection is indexed by attempt, so repeated valid outputs are retained and unsuccessful attempts contribute no training anchor. Candidates are filtered using a bounded quality score

$$\mathrm{Qual}(X_j, X^{\mathrm{cf}}) \in [0, 1]. \tag{A.3}$$

The score combines semantic consistency and boundary proximity. We use the product score

$$\mathrm{Qual}(X_j, X^{\mathrm{cf}}) = q_{\mathrm{sem}}(X_j, X^{\mathrm{cf}})\, q_{\mathrm{bd}}(X^{\mathrm{cf}}), \qquad q_{\mathrm{sem}}, q_{\mathrm{bd}} \in [0, 1]. \tag{A.4}$$

Semantic consistency. Let $e(\cdot)$ be a fixed embedding map, such as the penultimate-layer representation of the pre-unlearning model. For nonzero embeddings, semantic consistency is the rescaled cosine similarity; we set $q_{\mathrm{sem}} = 0$ when either embedding is zero:

$$q_{\mathrm{sem}}(X_j, X^{\mathrm{cf}}) = \frac{1 + \cos(e(X_j), e(X^{\mathrm{cf}}))}{2}. \tag{A.5}$$

Boundary proximity. Let $p_{\boldsymbol{\omega}}(x) = \mathrm{softmax}(h_{\boldsymbol{\omega}}(x))$ be the pre-unlearning predictive distribution. Let $p_{(1)}(x)$ and $p_{(2)}(x)$ denote its largest and second-largest coordinates, and define the probability margin

$$m(x) = p_{(1)}(x) - p_{(2)}(x). \tag{A.6}$$

Given a boundary bandwidth $\delta_{\mathrm{bd}} \in (0, 1]$, we set

$$q_{\mathrm{bd}}(X^{\mathrm{cf}}) = 1 - \min\left\{1, \frac{m(X^{\mathrm{cf}})}{\delta_{\mathrm{bd}}}\right\}. \tag{A.7}$$

Thus, q_{bd} is close to 1 when X^{cf} lies near the pre-unlearning decision boundary. With a threshold $\tau_q \in [0, 1]$, retain only feasible anchors satisfying

$$\widetilde{Z}_{\mathrm{cf}}(X_j) = \left\{X^{\mathrm{cf}} \in Z_{\mathrm{cf}}(X_j) : X^{\mathrm{cf}} \text{ feasible}, \quad \mathrm{Qual}(X_j, X^{\mathrm{cf}}) \geq \tau_q\right\}. \tag{A.8}$$

Teacher and student forward maps. Each network uses its own encoder and prediction head:

$$h_{\boldsymbol{\omega}} = g_{\boldsymbol{\omega}} \circ \phi_{\boldsymbol{\omega}}, \qquad h_{\boldsymbol{\omega}^u} = g_{\boldsymbol{\omega}^u} \circ \phi_{\boldsymbol{\omega}^u}. \tag{A.9}$$

For an input x, define

$$p_t(x) = \mathrm{softmax}(h_{\boldsymbol{\omega}}(x)), \qquad p_u(x) = \mathrm{softmax}(h_{\boldsymbol{\omega}^u}(x)). \tag{A.10}$$

Teacher outputs and generated inputs are detached during the student update, while gradients update both components of the student. The losses below use nonempty training minibatches $\mathcal{B}_r$ and $\mathcal{B}_f$.

Retain-side prediction alignment. Retained predictions are aligned on the same original input:

$$\mathcal{L}_r = \frac{1}{|\mathcal{B}_r|} \sum_{(X_j, Y_j) \in \mathcal{B}_r} \mathrm{KL}\big(p_t(X_j) \| p_u(X_j)\big). \tag{A.11}$$

Counterfactual anchor alignment. For each forget input, the teacher's counterfactual prediction is the target for the student's original-input prediction. Write $K_j = |\widetilde{Z}_{\mathrm{cf}}(X_j)|$ and set the empty anchor sum to zero:

$$\mathcal{L}_{\mathrm{cf}} = \frac{1}{|\mathcal{B}_f|} \sum_{X_j \in \mathcal{B}_f} \frac{1}{\max\{1, K_j\}} \sum_{X^{\mathrm{cf}} \in \widetilde{Z}_{\mathrm{cf}}(X_j)} \mathrm{KL}\big(p_t(X^{\mathrm{cf}}) \| p_u(X_j)\big). \tag{A.12}$$

At teacher initialization, differences between the original-input and counterfactual teacher predictions provide a redirection signal for the forget input. The deletion audit subsequently evaluates each resulting candidate.

Contrastive boundary stabilization. Use the student's original forget input and its counterfactual as the positive pair. Let $r_u(x) = \phi_{\boldsymbol{\omega}^u}(x) / \max\{\|\phi_{\boldsymbol{\omega}^u}(x)\|, \epsilon_{\mathrm{emb}}\}$ for fixed $\epsilon_{\mathrm{emb}} > 0$ and write $e_\tau(a, b) = \exp(\langle r_u(a), r_u(b) \rangle / \tau)$. A pre-specified training-only negative pool $\mathcal{N}_j(X^{\mathrm{cf}})$ excludes the teacher-predicted classes of both X_j and X^{cf}; pool construction and subsampling are independent of calibration observations. Then

$$\mathcal{L}_{\mathrm{cb}} = -\frac{1}{|\mathcal{B}_f|} \sum_{X_j \in \mathcal{B}_f} \frac{1}{\max\{1, K_j\}} \sum_{X^{\mathrm{cf}} \in \widetilde{Z}_{\mathrm{cf}}(X_j)} \log \frac{e_\tau(X_j, X^{\mathrm{cf}})}{e_\tau(X_j, X^{\mathrm{cf}}) + \sum_{X^- \in \mathcal{N}_j(X^{\mathrm{cf}})} e_\tau(X_j, X^-)}, \tag{A.13}$$

where $\tau > 0$ is fixed; an empty negative pool gives a zero summand. Gradients flow through all student branches. The negative-pool rule and the normalization above complete the predictive-alignment objective.

Overall objective. Define $\mathcal{L}_{\mathrm{sup}} = |\mathcal{B}_r|^{-1} \sum_{(X_i, Y_i) \in \mathcal{B}_r} \ell_{\mathrm{sur}}(h_{\boldsymbol{\omega}^u}(X_i), Y_i)$. For fixed nonnegative α, β, γ, the candidate objective is

$$\mathcal{L} = \mathcal{L}_r + \alpha \mathcal{L}_{\mathrm{cf}} + \beta \mathcal{L}_{\mathrm{cb}} + \gamma \mathcal{L}_{\mathrm{sup}}, \tag{A.14}$$

where ℓ_{sur} is a differentiable supervised surrogate loss. A gradient-descent step gives

$$\boldsymbol{\omega}^u \leftarrow \boldsymbol{\omega}^u - \eta \nabla_{\boldsymbol{\omega}^u} \mathcal{L}. \tag{A.15}$$

The protocol is controlled by

$$\lambda = [N_{\mathrm{cf}}, \tau_q, \alpha, \beta, \gamma, \eta], \tag{A.16}$$

which corresponds to the configuration vector described in the main text.

The exact one-step interface to Section 5 evaluates both directions at the same starting parameter θ:

$$g_f = \nabla(\mathcal{L}_r + \gamma \mathcal{L}_{\mathrm{sup}})(\theta), \qquad w_{\mathrm{cf}} g_{\mathrm{cf}} = \nabla(\alpha \mathcal{L}_{\mathrm{cf}} + \beta \mathcal{L}_{\mathrm{cb}})(\theta). \tag{A.17}$$

In this objective, the CF terms provide forget-input redirection, so their removal yields a retain-only update. The theorem compares the two directions at a common starting point; deployment comparisons also require both candidates to satisfy the same independent deletion audit. The same analysis can incorporate a base unlearner that supplies a deletion direction g_f. The weight w_{cf} scales the aggregate guidance direction and is distinct from the contrastive-loss coefficient β. With $N_{\mathrm{cf}} = 0$ or an empty feasible-anchor set, the CF losses are zero.

Algorithm 2 Counterfactual-guided candidate generation

Input: frozen teacher $h_{\boldsymbol{\omega}}$; fully trainable student initialized at $\boldsymbol{\omega}$; training sets $\mathcal{D}_f, \mathcal{D}_r$; fixed generator G_ϕ, feasibility and quality rules, negative-pool rule, configuration $\lambda = [N_{\mathrm{cf}}, \tau_q, \alpha, \beta, \gamma, \eta]$, temperature τ, and embedding floor ϵ_{emb}.
Output: post-unlearning model $h_{\boldsymbol{\omega}^u_\lambda}$.
for each training iteration **do**
  Sample minibatches $\mathcal{B}_f \subset \mathcal{D}_f$ and $\mathcal{B}_r \subset \mathcal{D}_r$.
  for each $X_j \in \mathcal{B}_f$ **do**
    $Z_{\mathrm{cf}}(X_j) \leftarrow G_\phi(X_j; N_{\mathrm{cf}})$.
    $\widetilde{Z}_{\mathrm{cf}}(X_j) \leftarrow \{X^{\mathrm{cf}} \in Z_{\mathrm{cf}}(X_j) : X^{\mathrm{cf}} \text{ feasible}, \ \mathrm{Qual}(X_j, X^{\mathrm{cf}}) \geq \tau_q\}$.
  end for
  Compute $\mathcal{L}_r$ on $\mathcal{B}_r$ using (A.11).
  Compute $\mathcal{L}_{\mathrm{cf}}$ using teacher(CF) targets and student(forget) inputs, Eq. (A.12); empty anchor sums are zero.
  Compute $\mathcal{L}_{\mathrm{cb}}$ using (A.13).
  Compute $\mathcal{L}_{\mathrm{sup}} \leftarrow |\mathcal{B}_r|^{-1} \sum_{(X_i,Y_i)\in\mathcal{B}_r} \ell_{\mathrm{sur}}(h_{\boldsymbol{\omega}^u}(X_i), Y_i)$.
  $\mathcal{L} \leftarrow \mathcal{L}_r + \alpha\mathcal{L}_{\mathrm{cf}} + \beta\mathcal{L}_{\mathrm{cb}} + \gamma\mathcal{L}_{\mathrm{sup}}$.
  $\boldsymbol{\omega}^u \leftarrow \boldsymbol{\omega}^u - \eta\nabla_{\boldsymbol{\omega}^u}\mathcal{L}$.
end for
return $h_{\boldsymbol{\omega}^u_\lambda}$.

B Proofs for Retain-Side Risk Control

All results in this section concern a bounded score $s_\lambda : \mathcal{Z} \to [0,1]$ and a calibration size $n = N_{\mathrm{cal}} \geq 1$. Let $\mathcal{F}$ contain the realized deletion request and all calibration-independent information that fixes the candidates, reference model, score, thresholds, and testing rule, including realized candidate-generation randomness. The sampling assumption is

$$(Z_1, \ldots, Z_n) \mid \mathcal{F} \sim P_r(\cdot \mid D_f)^{\otimes n}. \tag{B.1}$$

For almost every realization of $\mathcal{F}$, the candidates are fixed functions and their population risks are constants. Different candidates may have arbitrarily dependent scores on the same Z_i. The calibration design must provide this conditional i.i.d. sampling property in addition to data-set disjointness.

B.1 Proof of the fixed-configuration certificate

Proof of Proposition 4.1. Fix a realization of $\mathcal{F}$ and a candidate selected without inspecting the calibration sample. Write

$$R(\lambda) = \mathbb{E}_{Z\sim P_r(\cdot|D_f)}[s_\lambda(Z)], \qquad \widehat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} s_\lambda(Z_i). \tag{B.2}$$

Conditional on $\mathcal{F}$, the summands are i.i.d. bounded variables with mean $R(\lambda)$. For each fixed $a \in (0,1)$, under the unsafe null $R(\lambda) \geq a$, the Hoeffding–Bentkus statistic in Eq. (1) is a valid lower-tail p-value:

$$\Pr\Big\{B_n(\widehat{R}(\lambda), a) \leq t \mid \mathcal{F}\Big\} \leq t, \qquad 0 \leq t \leq 1. \tag{B.3}$$

Inverting these tests, with the empty inversion set assigned the value 1, gives

$$\Pr\{R(\lambda) \leq \widehat{a}_\lambda(\delta) \mid \mathcal{F}\} \geq 1 - \delta. \tag{B.4}$$

Integrating over $\mathcal{F}$ conditional on D_f proves the stated request-specific guarantee. □

The Bernoulli KL divergence in Eq. (1) uses its continuous extension: $0\log 0 = 0$, $h(0,a) = -\log(1-a)$, and $h(1,a) = -\log a$ for $0 < a < 1$. At $a = 1$ and $\widehat{r} < 1$, the lower-tail bound is 0; at $\widehat{r} = a = 1$, the $\widehat{r} \geq a$ branch is 1, so the UCB is 1. These conventions avoid undefined endpoint expressions. For calibration-based candidate selection, the simultaneous construction below provides numerical UCBs with post-selection coverage.

B.2 Proof of the valid configuration set

Proof of Proposition 4.2. Condition on $\mathcal{F}$. For a nonempty finite grid $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$, the unsafe hypotheses are

$$\mathcal{H}_j : R(\lambda_j) \geq r_0, \qquad 0 < r_0 < 1. \tag{B.5}$$

Their p-values are

$$p_j = \begin{cases} B_n(\widehat{R}(\lambda_j), r_0), & \widehat{R}(\lambda_j) < r_0, \\ 1, & \widehat{R}(\lambda_j) \geq r_0. \end{cases} \tag{B.6}$$

Each true-null p-value is super-uniform conditional on $\mathcal{F}$ by the preceding argument. Let $\mathcal{A}$ be a valid FWER procedure at level δ, such as Bonferroni or Holm, and set $\widehat{\Lambda}_{r_0} = \{\lambda_j : j \in \mathcal{A}(p_1, \ldots, p_N)\}$. Conditional FWER control gives

$$\Pr\Big\{\forall \lambda \in \widehat{\Lambda}_{r_0},\ R(\lambda) < r_0 \mid \mathcal{F}\Big\} \geq 1 - \delta. \tag{B.7}$$

This statement is also well-defined when no candidate is accepted. Integrating conditional on D_f proves the proposition. Intersecting the accepted set with a deletion-effective set can only remove candidates, so the same retain-side event covers every candidate in the intersection. □

Bonferroni and Holm allow arbitrary dependence between candidate p-values. An empty input grid returns the empty set. The intersection therefore combines the retain-side guarantee above with the evidence supplied by the chosen deletion audit.

The accepted set guarantees $R(\lambda) < r_0$ for every member, including a member selected after calibration. To report a numerical upper bound for that selected member, one can construct simultaneous UCBs. For $N \geq 1$, Bonferroni inversion gives

$$\widehat{a}_j^{\mathrm{sim}} = \widehat{a}_{\lambda_j}(\delta/N), \qquad \Pr\big\{\forall j,\ R(\lambda_j) \leq \widehat{a}_j^{\mathrm{sim}} \mid \mathcal{F}\big\} \geq 1 - \delta. \tag{B.8}$$

This follows by a union bound and remains true for any subsequently selected index. Threshold testing and simultaneous numerical inversion thus provide two complementary forms of retain-risk control.

C Proofs for Counterfactual Improvement

C.1 Details on the V_{gen}-generator assumption

Remark C.1 (Conditional generator-quality assumption). Let $\mathcal{G}$ fix the generator, quality function, threshold τ_q, and the input or input list before fresh generation seeds are drawn. It is a sub-σ-field of the non-calibration information $\mathcal{F}$, which additionally includes the realized seeds and resulting candidates. For a fixed input x, define

$$Q_x = q(x, G_\phi(x; \xi)) \in [0, 1]. \tag{C.1}$$

For a feasible output, q equals the declared quality score; failed and infeasible attempts receive quality zero. The yield result assumes $0 < \tau_q < \mu_q \leq 1$, making $Q_x \geq \tau_q$ exactly the event that an attempt passes feasibility and quality filtering. Both moments below are defined over all attempts, including failures. The yield analysis uses an interior threshold even though the candidate-generation protocol also allows $\tau_q \in \{0, 1\}$. The shared-input guarantee assumes, for every input to which it is applied,

$$\mathbb{E}[Q_x \mid \mathcal{G}] \geq \mu_q > \tau_q, \qquad \mathrm{Var}(Q_x \mid \mathcal{G}) \leq V_{\mathrm{gen}}^2, \tag{C.2}$$

with common constants μ_q, V_{gen}. The expectation and variance are over the fresh generator seed with x fixed. Independent seeds give conditional independence of repeated generations after the input has been fixed.

The gap $\mu_q - \tau_q$ measures expected quality above the filtering threshold, while V_{gen} controls generation variability. The resulting bound requires these conditional moment inequalities uniformly over the inputs considered. Monte Carlo estimates provide diagnostics for the moments.

C.2 Proof of the valid-anchor yield

Proposition C.2 (Conditional valid-anchor yield)**.** *Condition on $\mathcal{G}$ as above. Generate a fixed number $N_{\mathrm{cf}} \geq 1$ of anchors using independent seeds at one fixed input, or at a fixed list of inputs satisfying the same conditional moment bounds. Define*

$$I_t = \mathbf{1}\{Q_t \geq \tau_q\}, \qquad N_{\mathrm{valid}} = \sum_{t=1}^{N_{\mathrm{cf}}} I_t, \qquad \underline{p} = 1 - \frac{V_{\mathrm{gen}}^2}{V_{\mathrm{gen}}^2 + (\mu_q - \tau_q)^2}. \tag{C.3}$$

Then

$$\mathbb{E}[N_{\mathrm{valid}} \mid \mathcal{G}] \geq N_{\mathrm{cf}}\underline{p}. \tag{C.4}$$

For $0 < \nu < \underline{p}$,

$$\Pr\left\{\frac{N_{\mathrm{valid}}}{N_{\mathrm{cf}}} < \nu \mid \mathcal{G}\right\} \leq \exp[-2N_{\mathrm{cf}}(\underline{p} - \nu)^2]. \tag{C.5}$$

The same bound holds after averaging over $\mathcal{G}$ if the stated common conditional bounds hold almost surely.

Proof of Proposition C.2. Fix $\mathcal{G}$. For each input in the fixed generation schedule, Cantelli's inequality gives

$$\Pr(Q_t < \tau_q \mid \mathcal{G}) \leq \frac{\mathrm{Var}(Q_t \mid \mathcal{G})}{\mathrm{Var}(Q_t \mid \mathcal{G}) + (\mathbb{E}[Q_t \mid \mathcal{G}] - \tau_q)^2} \leq 1 - \underline{p}. \tag{C.6}$$

Thus $p_t = \mathbb{E}[I_t \mid \mathcal{G}] \geq \underline{p}$. Linearity of expectation yields the first conclusion without requiring independence. For the lower-tail conclusion, the independent seeds make the I_t conditionally independent, although different fixed inputs may give different success probabilities. Set $\bar{p} = N_{\mathrm{cf}}^{-1}\sum_t p_t \geq \underline{p}$. Hoeffding's inequality for independent bounded variables gives

$$\Pr\left\{\frac{1}{N_{\mathrm{cf}}}\sum_t I_t < \nu \mid \mathcal{G}\right\} \leq e^{-2N_{\mathrm{cf}}(\bar{p}-\nu)^2} \leq e^{-2N_{\mathrm{cf}}(\underline{p}-\nu)^2}. \tag{C.7}$$

Integrating proves the unconditional version. □

The generation budget and quality threshold are fixed before generation, and the lower-tail bound uses independent seeds. These conditions define the sampling design covered by the proposition; the valid-anchor fraction is used for $N_{\mathrm{cf}} \geq 1$.

C.3 Details on gradient alignment

Remark C.3 (Direction and magnitude of the alignment signal). For nonzero gradients define

$$G = \frac{\langle g_r(\theta), g_{\mathrm{cf}}\rangle}{\|g_r(\theta)\|\|g_{\mathrm{cf}}\|}. \tag{C.8}$$

Because the guided update subtracts g_{cf} with positive weight, $G > 0$ identifies a retain-surrogate descent direction at θ. The signal condition also requires sufficient gradient magnitude. One sufficient decomposition of Eq. (9) is

$$G \geq d^+ > 0, \qquad \|g_r(\theta)\|\|g_{\mathrm{cf}}\| \geq \kappa\nu > 0 \tag{C.9}$$

on $\mathcal{E}_{\mathrm{align}} \cap \mathcal{E}_{\mathrm{val}}$. This event requires both gradients to be nonzero. The parameter π_{align} lower-bounds the probability of the alignment event over non-calibration randomness. A union bound combines it with the valid-yield event and allows arbitrary dependence between the two events.

The factor $\kappa\nu$ is a lower bound on the product of gradient magnitudes. It is imposed on the aggregate guidance gradient with its declared normalization, accounting for possible cancellation between anchor gradients. It is an additional assumption beyond the valid-anchor count and gradient norm upper bounds.

C.4 PROOF OF CERTIFIED IMPROVEMENT

Proof of Theorem 5.1. Take $\eta > 0$ and $w_{\mathrm{cf}} > 0$, and assume the stated L-smoothness on a region containing the relevant update segments. Smoothness at θ^{std} gives

$$L_r(\theta^{\mathrm{std}}) - L_r(\theta^{\mathrm{cf}}) \geq \eta w_{\mathrm{cf}} \langle \nabla L_r(\theta^{\mathrm{std}}), g_{\mathrm{cf}} \rangle - \frac{L\eta^2 w_{\mathrm{cf}}^2}{2} \|g_{\mathrm{cf}}\|^2. \tag{C.10}$$

Since $\theta^{\mathrm{std}} = \theta - \eta g_f$, gradient Lipschitzness yields $\|\nabla L_r(\theta^{\mathrm{std}}) - g_r(\theta)\| \leq L\eta B_f$. By Cauchy–Schwarz,

$$L_r(\theta^{\mathrm{std}}) - L_r(\theta^{\mathrm{cf}}) \geq \eta w_{\mathrm{cf}} \langle g_r(\theta), g_{\mathrm{cf}} \rangle - L\eta^2 w_{\mathrm{cf}} B_f B_{\mathrm{cf}} - \frac{L\eta^2 w_{\mathrm{cf}}^2}{2} B_{\mathrm{cf}}^2. \tag{C.11}$$

On $\mathcal{E} = \mathcal{E}_{\mathrm{align}} \cap \mathcal{E}_{\mathrm{val}}$, the signal assumption makes this at least M_ν. Assume the common score-transfer margin $m_\nu = c_s M_\nu > 0$ of Eq. (13) holds on this event.

Now condition on the full non-calibration σ-field $\mathcal{F}$, including the realized candidate-generation randomness. The event $\mathcal{E}$ is $\mathcal{F}$-measurable and both candidates are fixed. For almost every realization in $\mathcal{E}$,

$$D_i = s_{\mathrm{std}}(Z_i) - s_{\mathrm{cf}}(Z_i) \in [-1, 1], \qquad \mathbb{E}[D_i \mid \mathcal{F}] \geq m_\nu, \tag{C.12}$$

and the D_i are conditionally i.i.d. Hence

$$\Pr\left\{ \frac{1}{n} \sum_i D_i \leq 0 \mid \mathcal{F} \right\} \leq e^{-n m_\nu^2/2} \quad \text{on } \mathcal{E}. \tag{C.13}$$

Taking conditional expectations gives

$$\Pr\left\{ \frac{1}{n} \sum_i D_i \leq 0,\ \mathcal{E} \right\} \leq e^{-n m_\nu^2/2} \Pr(\mathcal{E}) \leq e^{-n m_\nu^2/2}. \tag{C.14}$$

The HB inversion is nondecreasing in the empirical mean. Thus a positive paired empirical gap implies $\widehat{a}_{\mathrm{cf}}(\delta) \leq \widehat{a}_{\mathrm{std}}(\delta)$ when both inversions use the same n, sample, and δ. Adding $\Pr(\mathcal{E}^c) \leq (1 - \pi_{\mathrm{align}}) + \Pr(\mathcal{E}_{\mathrm{val}}^c)$ proves Eq. (14).

Each candidate is fixed with respect to calibration conditional on $\mathcal{F}$, so its marginal coverage follows independently from Proposition 4.1. At nominal level δ for each candidate, a union bound gives joint coverage at least $1 - 2\delta$; using $\delta/2$ for each gives at least $1 - \delta$. □

The proof first fixes the complete candidate realization, which gives conditionally independent paired calibration scores, and then averages over that realization. The common margin m_ν is a uniform positive lower bound throughout $\mathcal{E}$. The conclusion is a no-worse numerical certificate for the local one-step comparison.

For an explicit bounded-score transfer example, fix $M > 0$ before calibration and define

$$s_M(\theta, z) = \min\{1, [\ell_{\mathrm{sur}}(\theta; z) - \ell_{\mathrm{sur}}(\theta^{\mathrm{ref}}; z)]_+ / M\}. \tag{C.15}$$

This score is globally bounded. To obtain an explicit transfer relation, assume that the two candidates' pointwise loss differences relative to the same reference both lie strictly in $(0, M)$ for P_r-almost every z. Clipping is then inactive at both endpoints, so

$$\mathbb{E}[s_M(\theta^{\mathrm{std}}, Z) - s_M(\theta^{\mathrm{cf}}, Z)] = \frac{L_r(\theta^{\mathrm{std}}) - L_r(\theta^{\mathrm{cf}})}{M}. \tag{C.16}$$

The pointwise interior condition therefore instantiates Eq. (13) for this bounded surrogate score with $c_s = 1/M$. This condition concerns pointwise loss differences, rather than their average. Applying the comparison theorem to the default 0–1 score requires a score-transfer assumption for that score.

D SHIFT-AWARE CERTIFICATION AND ADDITIONAL RESULTS

D.1 PROOF OF THE DRIFT-AWARE CERTIFICATE

Proof of Theorem 6.1. Use the normalization $H^2(D,Q) = \frac{1}{2}\int(\sqrt{p_D}-\sqrt{p_Q})^2\,d\mu$ and $\mathrm{TV}(D,Q) = \frac{1}{2}\int|p_D - p_Q|\,d\mu$ for densities with respect to a common dominating measure. For a fixed bounded score,

$$\mathbb{E}_Q[s_{\mathrm{ret}}] - \mathbb{E}_D[s_{\mathrm{ret}}] \le \mathrm{TV}(D,Q) \le \sqrt{2}H(D,Q) \le \sqrt{2}\rho. \tag{D.1}$$

Conditional on the non-calibration information that fixes the candidate, its source-domain HB event has probability at least $1-\delta$ and implies $\mathbb{E}_D[s_{\mathrm{ret}}] \le \widehat{a}_D^{\mathrm{HB}}(\delta)$. Add the deterministic shift inequality, clip the resulting upper bound at 1, and integrate over the non-calibration information to obtain Eq. (15). □

For a finite-grid deployment target $R_Q(\lambda) < r_0$, the source unsafe null must be $R_D(\lambda) \ge r_0 - \sqrt{2}\rho$. When $r_0 > \sqrt{2}\rho$, applying a valid FWER procedure to these nulls gives a deployment-safe set by the same deterministic shift inequality. When $r_0 \le \sqrt{2}\rho$, this additive certificate returns the empty set. Alternatively, simultaneous source-domain UCBs may be used to retain only candidates satisfying $\widehat{a}_j^{\mathrm{sim}} + \sqrt{2}\rho < r_0$. The source threshold $r_0 - \sqrt{2}\rho$ reserves the required shift margin. For a calibration-selected candidate, a simultaneous numerical UCB supplies the corresponding drift headroom.

For completeness, an optional empirical-Bernstein base bound requires $n \ge 2$ and uses the unbiased variance

$$\widehat{V}_D = \frac{1}{n(n-1)}\sum_{i<j}(X_i - X_j)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{D.2}$$

Here X_i denotes the bounded score on the ith calibration example. The standard empirical-Bernstein term (Maurer & Pontil, 2009), followed by $+\sqrt{2}\rho$ and clipping, is another valid fixed-candidate shift bound. For $n = 1$, one can use HB inversion or the upper bound 1.

D.2 EFFECTIVE COUNTERFACTUAL RATE UNDER HELLINGER SHIFT

The next result complements retain-risk certification by bounding how a fixed generation mechanism's valid-anchor probability can change under distribution shift.

Proposition D.1 (Effective counterfactual rate under Hellinger shift)**.** *Let $H(D,Q) \le \rho$. Fix a common generator kernel $K(dz_{\mathrm{cf}} \mid z)$ and a measurable binary effectiveness rule $\pi : \mathcal{Z}\times\mathcal{Z} \to \{0,1\}$ for both distributions. Define*

$$f(z) = \int \pi(z, z_{\mathrm{cf}})K(dz_{\mathrm{cf}} \mid z), \qquad p_D = \mathbb{E}_D[f(Z)], \quad p_Q = \mathbb{E}_Q[f(Z)]. \tag{D.3}$$

Then $|p_Q - p_D| \le \sqrt{2}\rho$. If one input $Z \sim Q$ receives a fixed budget $N_{\mathrm{cf}} \ge 1$ of generation attempts, each with conditional marginal kernel K, and

$$N_{\mathrm{eff}} = \sum_{t=1}^{N_{\mathrm{cf}}} \pi(Z, Z_{\mathrm{cf},t}), \tag{D.4}$$

then

$$\mathbb{E}_Q[N_{\mathrm{eff}}] = N_{\mathrm{cf}}p_Q \ge N_{\mathrm{cf}}(p_D - \sqrt{2}\rho)_+. \tag{D.5}$$

Proof. Because π is binary, $f(z) \in [0,1]$ is the conditional probability that an anchor is effective. With the same function f under both distributions,

$$|p_Q - p_D| = |\mathbb{E}_Q f - \mathbb{E}_D f| \le \mathrm{TV}(D,Q) \le \sqrt{2}\rho. \tag{D.6}$$

For a fixed input, linearity of conditional expectation gives $\mathbb{E}[N_{\mathrm{eff}} \mid Z = z] = N_{\mathrm{cf}}f(z)$. Averaging over $Z \sim Q$ proves the expectation identity; nonnegativity of p_Q gives the positive-part lower bound. □

The expectation identity allows dependent anchors with the stated common conditional marginal kernel. Extending it to concentration bounds requires assumptions on that dependence. A soft effectiveness rule $\pi \in [0,1]$ gives the analogous bound for total effective weight. Changes to the generator or effectiveness rule between D and Q can be incorporated through a separate discrepancy term.

D.3 Comparing drift-aware certificates

The following result compares the optional empirical-Bernstein certificates through their empirical means and variances. The coverage levels for comparing or selecting candidates are given after the proof.

Proposition D.2 (Sufficient condition for a no-worse drift-aware certificate)**.** *Fix $\rho \geq 0$, $\delta \in (0,1)$, and $n \geq 2$. Consider two configurations evaluated on the same calibration sample. Let*

$$\widehat{a}_\lambda(\rho,\delta) = \min\{1, \widehat{R}_D(\lambda) + \kappa_\delta \sqrt{\widehat{V}_D(\lambda)} + c_{\rho,\delta}\}, \tag{D.7}$$

where

$$\kappa_\delta = \sqrt{\frac{2\ln(2/\delta)}{n}}, \qquad c_{\rho,\delta} = \frac{7\ln(2/\delta)}{3(n-1)} + \sqrt{2}\rho. \tag{D.8}$$

If, for some $\varepsilon \geq 0$,

$$\widehat{R}_D(\lambda_2) \leq \widehat{R}_D(\lambda_1) - \varepsilon, \quad \varepsilon \geq \kappa_\delta\big(\sqrt{\widehat{V}_D(\lambda_2)} - \sqrt{\widehat{V}_D(\lambda_1)}\big)_+, \tag{D.9}$$

then $\widehat{a}_{\lambda_2}(\rho,\delta) \leq \widehat{a}_{\lambda_1}(\rho,\delta)$.

Proof. Let $U(\lambda) = \widehat{R}_D(\lambda) + \kappa_\delta \sqrt{\widehat{V}_D(\lambda)} + c_{\rho,\delta}$. Subtracting the two expressions gives

$$U(\lambda_2) - U(\lambda_1) \leq -\varepsilon + \kappa_\delta\big(\sqrt{\widehat{V}_D(\lambda_2)} - \sqrt{\widehat{V}_D(\lambda_1)}\big). \tag{D.10}$$

If the variance difference is nonpositive, this is at most $-\varepsilon \leq 0$. Otherwise it is nonpositive by the assumed inequality. Monotonicity of $u \mapsto \min\{1,u\}$ preserves the ordering. □

For two fixed candidates, their nominal bounds have joint coverage at least $1-2\delta$ by a union bound. Selecting from a larger calibration-evaluated family requires a simultaneous construction such as Bonferroni bounds; the displayed algebra remains valid when the same adjusted level is used for both candidates.

E Additional Experiments

This appendix provides calibration-budget details (Appendix E.1), anchor-quality stress tests (Appendix E.2), architecture-transfer results and modality-specific implementation and analyses (Appendices E.3–E.5), and computational-cost and baseline comparisons (Appendices E.6–E.7).

Table 3: Configuration settings for the main vision unlearning and certification experiments.

Parameter	Setting
Datasets; backbone	CIFAR-100 and CUB-200; ResNet-18
Default retain split	Training/calibration/test: $80\%/10\%/10\%$
Vision CF construction	Headtails-style semantic-region replacement following (Chen et al., 2024)
Guided-update optimizer	Gradient descent; learning rate $\eta = 10^{-4}$; batch size 128
Random seed	42
Generation budget	$N_{\mathrm{cf}} \in \{0, 10, 50, 100\}$ generation attempts
Threshold and loss weights	$\tau_q, \alpha, \beta, \gamma$: default 0.2 each; one-at-a-time values $\{0.1, 0.2, 0.6, 1.0\}$
Candidate family	52 configurations per dataset/deletion setting
Table 1 CF budget	$N_{\mathrm{cf}} = 50$ for the 20-class deletion comparison
Default reference	Original pre-unlearning model
Fixed-candidate HB bound	$\delta = 0.1$ (90% confidence)
Family-wise selection	Bonferroni threshold $0.1/52 = 1/520$
Joint-screening tolerance	$r_0 = 0.05$ for the joint-set experiment
Deletion-audit thresholds	Class deletion: FA $\leq 10\%$; sample deletion: MIA advantage ≤ 0.05

For each generation budget, the default vector and three nondefault values for each of four parameters yield $1 + 4 \times 3 = 13$ distinct parameter vectors after deduplication, hence 52 across the four budgets. The definition of generation attempts is given in Appendix A.1.

E.1 Calibration-budget protocol

The calibration-budget comparison in Section 7.4 uses held-out retain-calibration sizes of $0.25\times$, $0.50\times$, $1.00\times$, and $2.00\times$ the default size. The HB inversion uses the empirical calibration score and sample size as inputs. At a fixed empirical calibration score, increasing the sample size tightens the Hoeffding component of the bound.

The retain-test risk is evaluated separately from calibration. Under deterministic evaluation of a fixed model on fixed test examples, this risk stays constant as calibration size changes. The small test-risk differences in the reported settings therefore reflect variation across those settings.

E.2 Stress test with noisy counterfactual anchors

Anchor quality and gradient alignment help explain the retain-side behavior of guided candidates. We test their sensitivity by deliberately degrading counterfactual anchors. Before quality filtering, we perturb generated anchors with Gaussian noise at levels 0, 0.02, 0.05, and 0.10, corresponding to clean, mild, moderate, and severe, respectively. We then measure the valid-anchor fraction, the cosine alignment between the retain gradient and the counterfactual guidance gradient, the retain-test degradation risk, and the certificate.

Figure 6 shows that noisy counterfactuals weaken the corrective signal. As noise increases, the valid-anchor fraction and gradient alignment decrease consistently. At the same time, the retain-test degradation risk increases and CMUCE outputs a looser certificate. For example, on CIFAR-100 class-level unlearning, the valid-anchor fraction decreases from 0.72 to 0.31, while the certificate increases from 0.050 to 0.079. The joint change in anchor quality, retain risk, and certificate tightness illustrates how CMUCE reflects the measured performance of its input candidates.

E.3 Transfer to transformer vision backbones

The main experiments use ResNet-18 as the default backbone. To test whether the certification behavior is tied to a convolutional architecture, we additionally evaluate ViT-Tiny and DeiT-Tiny, based on Vision Transformer (Dosovitskiy et al., 2021) and data-efficient image transformer (Touvron et al., 2021) architectures, while keeping the unlearning and certification protocol fixed.

Figure 7 shows that transformer backbones have slightly different degradation–utility trade-offs from ResNet-18. They often show slightly lower retain accuracy and slightly larger retain-side degradation risk or certificates. However, the overall certification behavior remains consistent. For example, on CIFAR-100 class-level unlearning, the certificate changes from 0.036 for ResNet-18 to 0.045

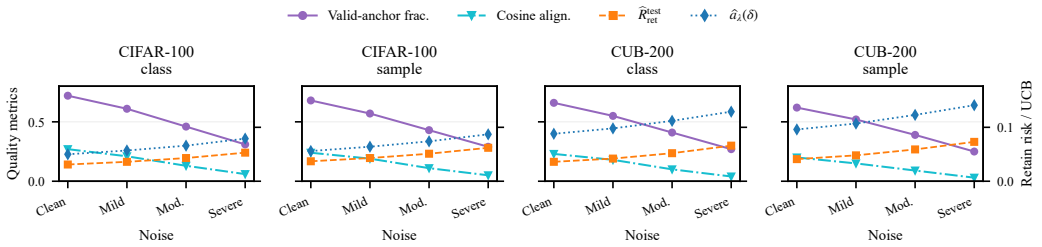

Figure 6: Stress test with noisy counterfactual anchors. Increasing anchor noise reduces valid-anchor fraction and gradient alignment, while increasing retain-test degradation risk and the certified upper bound.

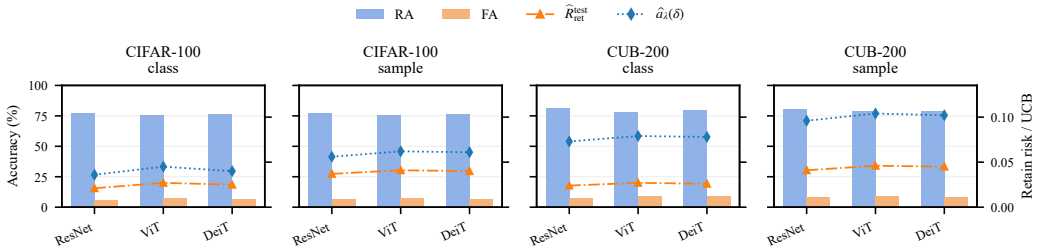

Figure 7: Backbone transfer analysis. CMUCE certification remains informative when replacing the ResNet-18 backbone with ViT-Tiny or DeiT-Tiny, although the absolute degradation–utility trade-off varies across architectures.

for ViT-Tiny and 0.040 for DeiT-Tiny, while the corresponding retain-test degradation risks remain close at 0.021, 0.027, and 0.025. The same certification procedure therefore applies across the tested convolutional and transformer architectures, with architecture-dependent degradation–utility trade-offs.

E.4 Implementation of the text and tabular pilots

This section details the text and tabular protocols used in Section 7.5. Each setting specifies an unlearning configuration, a bounded retain-side degradation score, and a held-out retain-calibration split independent of candidate construction.

For text, we use AG News (Zhang et al., 2015), a four-class news classification benchmark, and fine-tune DistilBERT (Sanh et al., 2019) as the base predictor. We consider two deletion regimes. The first is class-level deletion, where one news category is removed at a time. The second is sample-level deletion, where a fixed fraction of training examples is uniformly removed. To instantiate counterfactual guidance in text, we generate textual counterfactual anchors using FLAN-T5-small (Chung et al., 2024) as a lightweight rewriting model. Each anchor is filtered by a modality-specific quality function before being used in the unlearning objective. The quality function follows the same principle as in the vision setting, but uses text-appropriate criteria. In particular, it combines semantic consistency, which checks that the rewritten text remains close to the source content, and boundary proximity, which favors anchors near the model decision boundary and therefore provides a more informative correction signal for retain-side degradation control. The bounded retain-side degradation score is evaluated on the retain-calibration split to compute an HB upper bound. We also report degradation risk on the retain-test split.

For tabular data, we use Adult (Becker & Kohavi, 1996) and German (Hofmann, 1994) and train a three-layer DNN as the base predictor. Counterfactual anchors are generated using C-CHVAE (Pawelczyk et al., 2020). We use group-targeted sample deletion to create a non-uniform deletion request: the female group is identified from each dataset's sex information (jointly encoded with personal status in German), and 30% of its training samples request deletion. This removes selected records

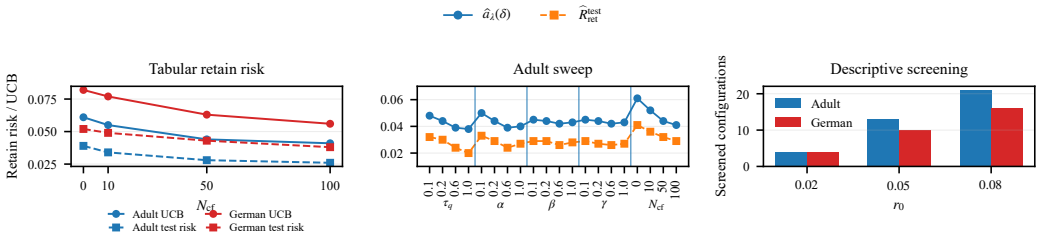

Figure 8: Additional tabular certification analyses. Left: retain risk and certificate as the generation budget N_{cf} increases on Adult and German. Middle: one-dimensional configuration dependence on Adult. Right: descriptive screening-set counts under different target retain-risk tolerances.

while retaining the feature schema. The resulting group imbalance provides a stress test for retain-side degradation control. We report the retain-test degradation risk and its HB certificate computed on the held-out retain-calibration split.

Across these modalities, the certification conditions are the same: a bounded pre-fixed score, conditional i.i.d. calibration independent of candidate construction, and the appropriate multiplicity correction.

E.5 Additional tabular certification analyses

We examine fixed-candidate certificates, configuration dependence, and descriptive screening-set counts on tabular data. The goal is to check whether the qualitative behavior of CMUCE is preserved when the data modality, predictor class, and counterfactual generator all differ from the vision setup.

The left panel of Figure 8 studies how retain-test degradation risk and the HB upper bound change as the counterfactual generation budget N_{cf} increases. We report results for both Adult and German. The main pattern is the same as in the vision experiments. As the generation budget increases, the observed retain-test degradation risk decreases and the certificate tightens. This pattern associates broader counterfactual generation with improved retain performance and tighter certificates in the reported settings.

The middle panel studies one-dimensional configuration dependence on Adult. We vary each configuration component while holding the remaining components fixed, following the same protocol used in the main vision experiments. The reported curves show how the retain-side test degradation risk and the certificate respond to these changes. The qualitative trend again matches the vision setting. Extremely weak or extremely aggressive choices can worsen either utility or certified degradation behavior, while moderate values produce more stable results. The same evaluation can therefore help identify useful configuration ranges for tabular predictors.

The right panel shows descriptive screening-set counts as the target retain-risk tolerance varies. For formal configuration selection on a specified tabular grid Λ, the protocol in Appendix B.2 tests all $N = |\Lambda|$ unsafe nulls with the Bonferroni threshold δ/N. The family size is set by that tabular grid.

E.6 Efficiency overhead

The evaluated pipeline combines counterfactual-guided candidate generation with retain-risk certification. We separate total runtime into the base update, counterfactual generation, and calibration pass, and also report peak GPU memory.

Figure 9 shows that the main computational cost comes from the base unlearning update, followed by counterfactual generation. The calibration pass is comparatively inexpensive: it evaluates a bounded score on the retain-calibration split and computes a scalar HB certificate. For example, on CIFAR-100 class-level unlearning with $N_{\mathrm{cf}} = 50$, the total time is 53.2 seconds, of which 42.1 seconds come from base unlearning, 8.4 seconds from counterfactual generation, and 2.7 seconds from calibration. The peak GPU memory remains within a few GB in all reported settings.

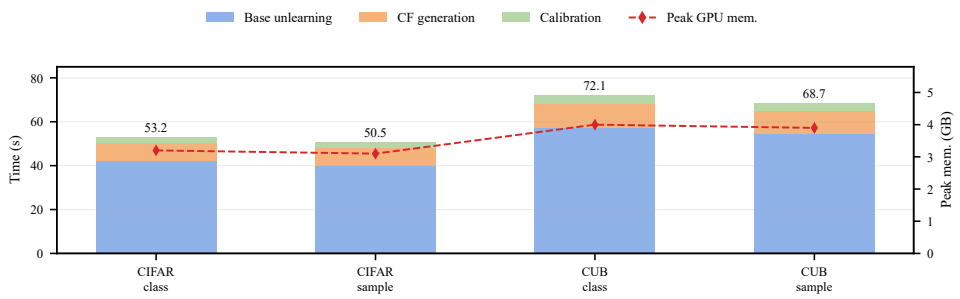

Figure 9: Efficiency breakdown of CMUCE. Stacked bars show runtime components, and the dashed line reports peak GPU memory.

E.7 Certified comparison with non-counterfactual baselines

We apply the same HB retain-risk certification procedure to CF-guided candidates, FT-Unlearn, and Scrub. This common evaluation reports both observed degradation and a certified upper bound for each candidate generator. The labels FT-Unlearn+HB and Scrub+HB denote the corresponding certified baselines. Figure 10 shows lower retain-test degradation risks and tighter UCBs for the CF-guided candidates in the reported settings. All methods share the bounded score and calibration procedure, illustrating how the reusable certificate reflects candidate quality.

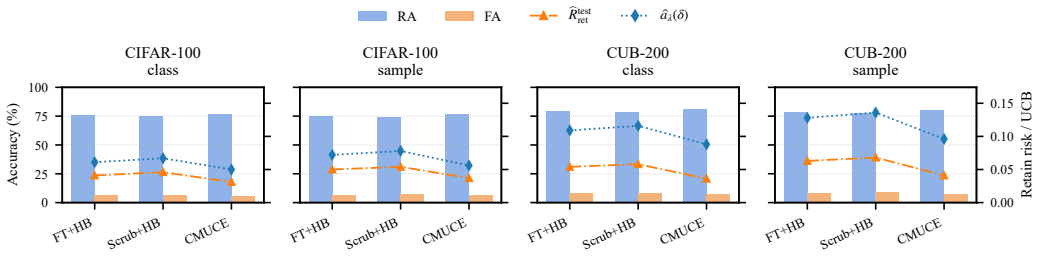

Figure 10: Certified comparison with non-counterfactual baselines. All methods use the same bounded retain-side degradation score and HB certification procedure. The CF-guided candidates have lower empirical degradation risks and tighter certificates in the reported settings.