RETAIN-UNLEARN GATED GRADIENT UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose TVE-Fisher, a simple and principled approach to unlearning in multimodal models that makes task-arithmetic editing selective. We cast unlearning as a retention-constrained local edit and observe an asymmetry of orders: forgetting grows linearly with the edit, whereas retention damage has no first-order term and is governed by Fisher information (FI). Retain curvature should therefore modulate the forget gradient. Under a diagonal geometry, this yields an optimal parameter-selection score, which we relax into an order-preserving soft mask with guarantees on optimality and robustness to estimation error. Concretely, TVE-Fisher computes retain FI statistics once, on a subset of observable retain data under the transductive protocol, to identify parameters critical to retained knowledge. It tracks forget FI statistics dynamically with an exponential moving average to capture the parameters currently relied upon for fitting the forget data. The resulting mask gates gradient updates during forget-side fine-tuning. We then extract a task vector from the difference between the initial and fine-tuned parameters and negate it; the same analysis shows that negation preserves the selected parameters and the retention cost while reversing the forget effect. Extensive experiments on standard vision–language model (VLM) and multimodal large language model (MLLM) unlearning benchmarks across multiple backbones show that TVE-Fisher achieves a stronger forgetting–retention trade-off than related baselines.

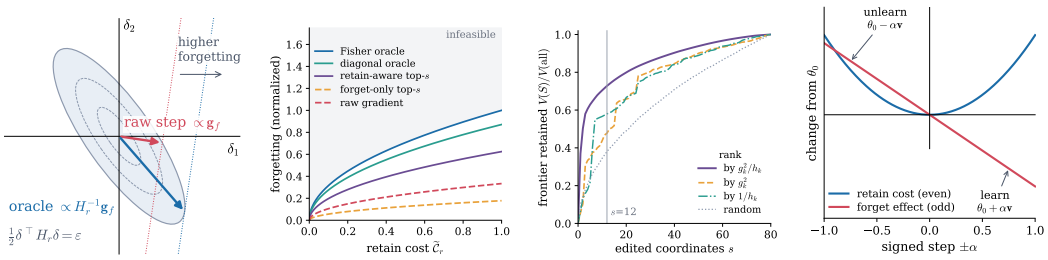
1 INTRODUCTION

Vision–language models (VLMs) have become a central paradigm for transferable representation learning, owing to large-scale image–text pre-training and strong zero-shot generalization Radford et al. (2021); Jia et al. (2021); Li et al. (2021); Alayrac et al. (2022); Chen et al. (2022); Li et al. (2023). Because they are trained on broad, heterogeneous, and weakly curated web-scale data, it is increasingly difficult to control what they contain Schuhmann et al. (2021; 2022); Gadre et al. (2023). In realistic deployments, one may need to remove the influence of privacy-sensitive, harmful, proprietary, or otherwise non-compliant data while preserving the model's remaining capabilities European Union (2016); Ginart et al. (2019); Guo et al. (2020); Bourtoule et al. (2021); Chundawat et al. (2023); Wang et al. (2024b); Nguyen et al. (2025); Dong et al. (2025). This is particularly challenging in multimodal systems, where the goal is often to forget a specific visual capability or task behavior without eroding general-purpose visual knowledge Yang et al. (2025); Kravets & Namboodiri (2025); Ma et al. (2025), and to do so efficiently, without retraining the full model.

A natural starting point is *task arithmetic*, in which a task vector is the parameter difference between a pre-trained model and its fine-tuned version Ilharco et al. (2023). This view is closely related to weight-space interpolation, model averaging, and model editing Wortsman et al. (2022); Matena & Raffel (2022); Yadav et al. (2023). Adding a task vector transfers a behavior, and negating it suppresses that behavior Ilharco et al. (2023); Dong et al. (2025). For large VLMs this is attractive: it is lightweight, adds no inference-time overhead, and avoids retraining. Yet standard task arithmetic edits all parameters along a single direction with a uniform strength, even though parameters differ both in how much they contribute to forgetting and in how much their change harms retained capabilities. This raises a central question: *how should a task-vector edit be distributed across parameters?*

We answer this question by treating unlearning as a retention-constrained local edit, and the answer rests on an asymmetry of orders. Forgetting grows linearly with the edit and is governed by the forget

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.



(a) Retain curvature reshapes the forget direction (b) Damped Pareto frontier (c) Which coordinates to edit (d) Negation: even cost, odd effect

Figure 1: Motivation for TVE-Fisher: retain curvature (second order) modulates the forget gradient (first order), yielding edits closer to the forget–retain Pareto frontier.

gradient, whereas retention damage, measured as the KL divergence from the original predictions, has no first-order term and is governed by the Fisher information Fisher (1925); Amari (1998); Martens & Grosse (2015). Retain curvature should therefore *modulate* the forget signal (Fig. 1). Fig. 1a shows that, under the same retention budget, the Fisher-preconditioned edit reaches higher forgetting than the raw forget gradient. Fig. 1b shows that forget-only parameter selection can fall even below the raw gradient, because it edits exactly the parameters retention relies on, while retain-aware selection stays close to the optimum. Fig. 1c shows that ranking parameters by the squared forget gradient over the retain Fisher recovers most of the achievable forgetting with few edits, far more efficiently than either signal alone or random selection. Fig. 1d shows that the retention cost is even in the edit while the forget effect is odd, so negating a learned task vector reverses forgetting at the same retention cost. Fisher information thus enters as the leading term of retention damage, not merely as an importance heuristic Kirkpatrick et al. (2017); Aljundi et al. (2018); Zenke et al. (2017); Benzing (2022); Matena & Raffel (2022).

Building on this, we propose **TVE-Fisher**, a Fisher-guided task-vector unlearning framework that acts as a filtered version of task arithmetic. Fisher statistics are converted into a soft mask that gates gradient updates during forget-side fine-tuning. We derive this mask as an order-preserving relaxation of the optimal parameter-selection rule, so every threshold of the mask recovers an optimal set of edited parameters, and selection degrades gracefully under estimation error. A visual task vector is then extracted and negated Ilharco et al. (2023). The same local analysis justifies negation: the selected parameters do not depend on the sign of the forget gradient, and the retention cost is symmetric under negation while the forget effect reverses. Experiments with CLIP backbones on eight vision benchmarks show that TVE-Fisher, while as lightweight as standard task arithmetic, consistently improves the forgetting–retention trade-off over it and over several task-vector-based baselines. Layer- and parameter-level analyses further show that Fisher information gives an interpretable account of which parts of the model drive forgetting and which support retention. Our main contributions are summarized as follows:

i. We formulate VLM unlearning as a retention-constrained local edit and show that retain curvature should modulate the forget gradient. This yields a forget–retain frontier on which any editing direction is characterized by a single efficiency measure.
ii. We derive the optimal parameter-selection score, the squared forget gradient over the retain Fisher, and relax it into an order-preserving soft mask with guarantees on optimality and robustness to estimation error.
iii. We instantiate this principle as TVE-Fisher, which combines a static retain-side Fisher with a dynamically tracked forget-side Fisher, and we show why task-vector negation preserves both the selected parameters and the retention cost.
iv. We provide layer- and parameter-level analyses and ablations over three masking forms, showing how Fisher-guided masks reveal the parts of the model most responsible for forgetting and retention.

2 Related Work

Unlearning for vision–language models. Vision–language pre-training has evolved from early models such as ViLBERT, VisualBERT, UNITER, and ViLT to stronger web-scale systems including

CLIP, ALIGN, ALBEF, SimVLM, BLIP-2, and CoCa Lu et al. (2019); Li et al. (2019); Chen et al. (2020); Kim et al. (2021); Radford et al. (2021); Jia et al. (2021); Li et al. (2021); Wang et al. (2021); Li et al. (2023); Yu et al. (2022). As these models absorb broad and weakly curated multimodal data, removing unwanted knowledge without retraining has become increasingly important and difficult. Compared with unimodal unlearning Ginart et al. (2019); Guo et al. (2020); Bourtoule et al. (2021); Chundawat et al. (2023), VLM unlearning remains relatively underexplored and is complicated by cross-modal entanglement. Recent work includes task-vector-style unlearning Dong et al. (2025), class/concept removal and synthetic-forget settings Kravets & Namboodiri (2025); Li et al. (2024); Geng & Li (2025), association-level editing in CLIP Yang et al. (2025), privacy-oriented evaluation benchmarks Ma et al. (2025); Dontsov et al. (2025), and domain- or robustness-driven formulations Kawamura et al. (2025); Han et al. (2025). The present work focuses on making task-vector unlearning itself more selective by identifying which parameters should be updated for erasure of data, and which should remain protected.

Task vector arithmetic and model editing. Task arithmetic represents task-specific behavior as the difference between a fine-tuned checkpoint and its base model, enabling addition or negation of task directions for lightweight editing Ilharco et al. (2023). This idea is closely tied to weight interpolation and model merging, including model soups, Fisher-weighted averaging, AdaMerging, and TIES-Merging Wortsman et al. (2022); Matena & Raffel (2022); Yang et al. (2024); Yadav et al. (2023). For unlearning, it has also been combined with function-level ensembling and simplex-style constructions Dong et al. (2025). NegMerge Kim et al. (2025) merges task vectors from different fine-tuning configurations by retaining sign-consistent coordinates before negation. OrthoReg Liu et al. (2026) promotes weight disentanglement by regularizing the internal orthogonality of weight updates during fine-tuning. Recent extensions further study tangent-space arithmetic, sparse task-localized editing, layer-structured task vectors, coefficient optimization, and theoretical conditions for successful editing Ortiz-Jimenez et al. (2023); Iurada et al. (2025); Gargiulo et al. (2025); Yoshida et al. (2025); Li et al. (2025). Our method broadly fits into this task-vector editing family while emphasizing selective construction of the unlearning direction.

Fisher information and parameter importance. Fisher information is a classical measure of parameter informativeness in statistical estimation Fisher (1925). In deep learning, related importance signals have been widely used to protect parameters in continual learning, including EWC and online EWC, Synaptic Intelligence, MAS, RWalk, Progress & Compress, and variational continual learning Kirkpatrick et al. (2017); Ritter et al. (2018); Zenke et al. (2017); Aljundi et al. (2018); Chaudhry et al. (2018); Schwarz et al. (2018); Nguyen et al. (2018). Broader analyses connect these regularizers to shared notions of sensitivity and posterior precision Benzing (2022); Wang et al. (2024a); Daxberger et al. (2021), while related ideas have also appeared in incremental learning without data replay Dhar et al. (2019). Fisher-weighted model merging extends the same intuition to checkpoint combination Matena & Raffel (2022). These lines together motivate Fisher information diagonal as a practical parameter-wise signal for selective updates in model editing and unlearning.

3 METHODOLOGY

We cast selective task erasure as a retention-constrained local editing problem. Retention damage is locally second order in the edit, whereas the forgetting objective changes at first order, suggesting that retain-Fisher curvature should modulate the forget signal (Fig. 1). Our derivation starts from the resulting full-space oracle, then reduces it to a diagonal coordinate score and a soft selection rule, replaces the inaccessible quantities with empirical statistics that are practical at VLM scale, and finally constructs the unlearning edit through task-vector negation. The oracle and coordinate-selection results are exact for the damped local surrogate; the empirical mask and deployment procedure are scalable instantiations whose approximations are stated explicitly. Detailed derivations and proofs are deferred to Appendix A.

3.1 RETENTION-CONSTRAINED ORACLE EDIT

Setting. Let $\boldsymbol{\theta}_0$ be a pretrained VLM with its text encoder frozen, and let $\boldsymbol{\delta}$ be an edit to the visual parameters. For forget data $\mathcal{D}_f$ and retain data $\mathcal{D}_r$, we seek large forgetting utility $\mathcal{U}_f$ at small change of the original predictions. We take $\mathcal{U}_f(\boldsymbol{\delta}) = \ell_f(\boldsymbol{\theta}_0 + \boldsymbol{\delta}) - \ell_f(\boldsymbol{\theta}_0)$, the increase of the forget loss, so that $\mathbf{g}_f = \nabla \ell_f(\boldsymbol{\theta}_0)$ and learning the forget task moves along $-\mathbf{g}_f$. We use a budget rather than a penalty because, by Lagrange duality, the two are equivalent locally, and

varying the budget traces the entire trade-off. Writing $\mathbf{g}_f = \nabla_{\boldsymbol{\delta}} \mathcal{U}_f(\mathbf{0})$, forgetting is first order, $\mathcal{U}_f(\boldsymbol{\delta}) = \mathcal{U}_f(\mathbf{0}) + \mathbf{g}_f^\top \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^2)$. Retention damage is

$$\mathcal{C}_r(\boldsymbol{\delta}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_r}[\mathrm{KL}(p_{\boldsymbol{\theta}_0}(\cdot \mid \mathbf{x}) \,\|\, p_{\boldsymbol{\theta}_0+\boldsymbol{\delta}}(\cdot \mid \mathbf{x}))] = \tfrac{1}{2}\boldsymbol{\delta}^\top F_r \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3), \tag{1}$$

whose zeroth- and first-order terms vanish because the KL is minimized at $\boldsymbol{\theta}_0$:

$$\nabla \mathcal{C}_r(\mathbf{0}) = -\mathbb{E}_{\mathbf{x}} \textstyle\sum_y p_{\boldsymbol{\theta}_0}(y \mid \mathbf{x}) \nabla \log p_{\boldsymbol{\theta}_0}(y \mid \mathbf{x}) = -\mathbb{E}_{\mathbf{x}} \nabla \sum_y p_{\boldsymbol{\theta}_0}(y \mid \mathbf{x}) = \mathbf{0}. \tag{2}$$

The leading term of the retention damage is therefore the predictive Fisher F_r. This asymmetry is the central observation of the paper. Forgetting grows linearly in the edit, while retention damage grows only quadratically, with curvature F_r. Small edits aligned with low-curvature directions are therefore cheap. The Fisher enters as the first nonvanishing term of the damage, not as a heuristic measure of parameter importance. Because F_r may be singular, we work with the damped surrogate $\widetilde{\mathcal{C}}_r(\boldsymbol{\delta}) = \frac{1}{2}\boldsymbol{\delta}^\top H_r \boldsymbol{\delta}$, where $H_r = F_r + \gamma I$ and $\gamma > 0$. This surrogate is conservative: it upper-bounds the quadratic term of (1).

Oracle and frontier. For a retention budget ε, the surrogate problem and its solution are

$$\max_{\boldsymbol{\delta}} \ \mathbf{g}_f^\top \boldsymbol{\delta} \ \text{ s.t. } \ \tfrac{1}{2}\boldsymbol{\delta}^\top H_r \boldsymbol{\delta} \le \varepsilon, \qquad \boldsymbol{\delta}_{\mathrm{orc}} = \sqrt{\frac{2\varepsilon}{\mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f}}\, H_r^{-1} \mathbf{g}_f. \tag{3}$$

The retain curvature does not merely flag "important parameters". It rotates and rescales the forget direction, and the attained utility $U^\star(\varepsilon) = \sqrt{2\varepsilon\, \mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f}$ traces the forget–retain Pareto frontier of the surrogate.

Proposition 1 (Frontier efficiency) *For any direction* $\mathbf{v}$*, the edits* $\alpha\mathbf{v}$ *with* $\alpha \ge 0$ *attain the curve* $c(\mathbf{v})\, U^\star(\varepsilon)$*, where*

$$c(\mathbf{v}) = \left|\cos_{H_r}\!\left(\mathbf{v},\, H_r^{-1}\mathbf{g}_f\right)\right| \le 1, \qquad c(\mathbf{g}_f) \ge \frac{2}{\sqrt{\kappa} + 1/\sqrt{\kappa}}, \quad \kappa = \frac{\lambda_{\max}(H_r)}{\lambda_{\min}(H_r)}. \tag{4}$$

Locally, every editing method is therefore characterized by a single cosine in the retain metric. Its forget–retain curve is the frontier scaled by $c(\mathbf{v})$, and the deployment coefficient α moves along that curve. This gives a common yardstick for unlearning directions that compares whole curves rather than single operating points. The frontier is concave with slope $\propto \varepsilon^{-1/2}$, so the first units of forgetting are the cheapest, and the height of the frontier, $\mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f$, measures the local compatibility of the forget and retain objectives. The oracle itself is infeasible, since a VLM-scale H_r cannot be stored or inverted.

3.2 From the oracle to a coordinate score

Diagonal geometry. Replacing H_r by $D_r = \mathrm{diag}(H_r)$ decouples the coordinates:

$$\boldsymbol{\delta}_{\mathrm{diag}} \propto D_r^{-1} \mathbf{g}_f, \qquad (\delta_{\mathrm{diag}})_k \propto \frac{g_{f,k}}{h_k}, \quad h_k = (F_r)_{kk} + \gamma. \tag{5}$$

Each coordinate is judged by a forget benefit divided by a retain cost. In the D_r-norm, the relative error with respect to $H_r^{-1}\mathbf{g}_f$ is at most $\chi_r = \max\{1 - 1/\lambda_{\max}(R_r),\, 1/\lambda_{\min}(R_r) - 1\}$, where $R_r = D_r^{-1/2} H_r D_r^{-1/2}$ (Lemma 1). The approximation is therefore accurate when the correlation-normalized retain geometry is close to the identity, and the bound requires only $H_r \succ 0$. For a trainable subspace $\boldsymbol{\delta} = M\mathbf{c}$, including the tangent space of a rank-one booster, the same derivation applies with $\mathbf{g}_f \to M^\top \mathbf{g}_f$ and $H_r \to M^\top H_r M$ (App. A.4).

Optimal support. To expose the selection structure, consider an editor that may modify only a support S of size s.

Proposition 2 (Optimal support) *Under the diagonal metric, the best edit supported on* S *attains* $V(S)^2 = 2\varepsilon \sum_{k\in S} \sigma_k$*, with* $\sigma_k = g_{f,k}^2/h_k$*. The optimal support of size* s *is therefore the set of the* s *largest scores* σ_k*.*

The forget gradient enters squared because σ_k is a contribution to $V(S)^2$, not a step size. No forget curvature is involved: σ_k is first-order forget information modulated by second-order retain curvature. The selected support also retains a fraction $(\sum_{k\in S} \sigma_k / \sum_k \sigma_k)^{1/2}$ of the diagonal frontier.

How quickly this fraction saturates in s measures how localized the forget task is in parameter space. When forget sensitivity and retain curvature are correlated, the forget-only rule concentrates the edit on exactly the coordinates retention needs. Its frontier can then fall below that of the raw gradient (Fig. 1).

Relaxation to a soft mask. Let $\mathbf{z} \in \{0,1\}^d$ indicate the support. The hard top-s rule is discontinuous in the scores and brittle under estimation noise, so we relax it. Because $V(S)^2$ is modular, the relaxation is unusually clean.

Proposition 3 (Soft selection) *(i) Relaxing $\mathbf{z}$ to $[0,1]^d$ is exact: the linear program has the top-s indicator as an optimum. A fractional z_k is an inclusion probability, and $\mathbb{E}\, V(S)^2 = 2\varepsilon\, \boldsymbol{\sigma}^\top \mathbf{z}$.*
(ii) Adding $-\frac{1}{2\beta}\|\mathbf{z}\|_2^2$ yields $z_k = \mathrm{clip}(\beta(\sigma_k - \nu), 0, 1)$. This loses at most $s/(2\beta)$ in objective value, tends to top-s as $\beta \to \infty$, and contains the proportional mask $\mathbf{z} \propto \boldsymbol{\sigma}$ as the member with $\nu = 0$.
(iii) Every such mask is nondecreasing in σ_k, so each of its superlevel sets is, up to ties, an optimal support of its own size.
(iv) If estimated scores satisfy $a \le \hat{\sigma}_k/\sigma_k \le b$ with $\sigma_k > 0$, the selected support retains at least $\sqrt{a/b}$ of the optimal frontier value.

A soft mask therefore encodes the whole nested family of optimal supports at once: thresholding it at any level recovers an exact discrete optimum. Part (iv) turns every estimation choice below into an explicit multiplicative distortion (proofs in App. A.6–A.7).

3.3 TVE-Fisher

Empirical statistics. We estimate the score σ_k from squared per-example gradients,

$$\boldsymbol{\zeta}(\boldsymbol{\theta}; \mathcal{D}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\nabla_{\boldsymbol{\theta}^{(v)}} \ell \odot \nabla_{\boldsymbol{\theta}^{(v)}} \ell \right]. \tag{6}$$

The retain statistic $\boldsymbol{\zeta}^{(r)} = \boldsymbol{\zeta}(\boldsymbol{\theta}_0; \mathcal{D}_r)$ is computed once, with pseudo-labels from the frozen model, as a proxy for $\mathrm{diag}(F_r)$. Sampling labels from $p_{\boldsymbol{\theta}_0}$ would make it unbiased; hard pseudo-labels give an empirical-Fisher proxy (Kunstner et al., 2019). The forget statistic $\widehat{\boldsymbol{\zeta}}_t^{(f)} = \boldsymbol{\zeta}(\boldsymbol{\theta}_t; \mathcal{D}_f)$ is recomputed each epoch with ground-truth labels as a proxy for $g_{f,k}^2$; it is upward-biased by a factor $1 + 1/\mathrm{SNR}_k$. Retention is thus anchored at the reference model, while the forget signal tracks what remains forgettable. We smooth the forget statistic by

$$\bar{\boldsymbol{\zeta}}_t^{(f)} = \mu \bar{\boldsymbol{\zeta}}_{t-1}^{(f)} + (1-\mu)\widehat{\boldsymbol{\zeta}}_t^{(f)}. \tag{7}$$

Calibration and mask. Each epoch, both statistics are mapped by one shared affine map,

$$\mathrm{cal}_t(\mathbf{z}) = \frac{\mathbf{z} - z_{\min}}{z_{\max} - z_{\min} + \epsilon}, \qquad z_{\min}, z_{\max} = \min, \max\Big([\bar{\boldsymbol{\zeta}}_t^{(f)}; \boldsymbol{\zeta}^{(r)}] \Big), \tag{8}$$

and we reuse the symbols for the calibrated copies. TVE-Fisher then uses

$$\mathbf{m}_t = \frac{\bar{\boldsymbol{\zeta}}_t^{(f)}}{(1 + \boldsymbol{\zeta}^{(r)})^\lambda} \qquad \textbf{(Mask iii)}. \tag{9}$$

Expanding the calibration shows that (9) has the tempered form of the score σ_k, up to a global factor and a numerator shift by $z_{\min}$. Its effective retain denominator lies in $[R + \epsilon,\ 2R + \epsilon]$ with $R = z_{\max} - z_{\min}$, so its dynamic range is below 2 (App. A.8). At $\lambda = 1$, Mask (iii) is the proportional member of Proposition 3. Other values of λ temper retain protection, at a cost of at most a factor $2^{-|1-\lambda|/2}$ in frontier value. We ablate two alternatives:

$$\mathbf{m}_t^{(i)} = \frac{\bar{\boldsymbol{\zeta}}_t^{(f)}}{1 + \lambda \bar{\boldsymbol{\zeta}}_t^{(f)} \odot \boldsymbol{\zeta}^{(r)}} \qquad \textbf{(Mask i)}, \tag{10}$$

$$\mathbf{m}_t^{(ii)} = \bar{\boldsymbol{\zeta}}_t^{(f)} \odot \max(0, 1 - \lambda \boldsymbol{\zeta}^{(r)}) \qquad \textbf{(Mask ii)}. \tag{11}$$

Mask (i) is a stabilized SGD-style surrogate and Mask (ii) is the clipped linearization of (9). Neither preserves the ordering of σ_k in general.

Interaction with AdamW. The mask multiplies gradients before AdamW normalization (Kingma & Ba, 2014). Under an idealized stationary second-moment model, a constant positive gate cancels,

so the gate acts mainly as coordinate selection rather than as a per-coordinate step size. This is precisely the regime described by Propositions 2 and 3, which is why we derive the mask as a selection rule rather than as an allocation of step sizes. The implicit threshold of AdamW is a level set of $m_k\sqrt{v_k}$ rather than of m_k. Decoupled weight decay still moves gated-out coordinates unless they are excluded from it (App. A.9).

When does retention add information? If the calibrated statistics are proportional, $\bar{\boldsymbol{\zeta}}_t^{(f)} = c\boldsymbol{\zeta}^{(r)}$, then Mask (iii) reduces to a fixed scalar function of the forget statistic, and the retain signal adds no independent coordinate information. For $\lambda \leq 1$, it then also leaves the ranking unchanged (App. A.11). TVE-Fisher is therefore most useful when forget sensitivity and retain curvature have different coordinate structure. Gains over a forget-only gate should be read together with the disagreement between the two statistics.

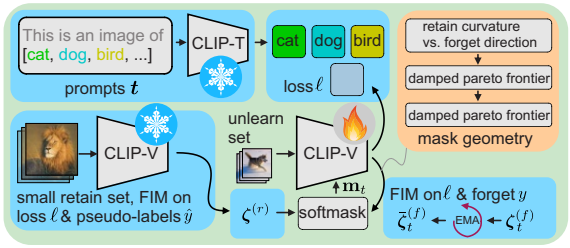


Figure 2: Our TVE-Fisher-FFT pipeline. On small retain set, we obtain pseudo-labels $\hat{y}$ and retain FIM vector $\boldsymbol{\zeta}^{(r)}$ from the frozen CLIP (vision). We obtain forget $\boldsymbol{\zeta}_t^{(f)}$ per epoch on unlearn set using Exponential Moving Average (EMA), and calculate soft mask $\mathbf{m}_t$ gate for gradients based on mask geometry analysis in Fig. 1. We fine-tune the model for $t = 1, \ldots, T$.

3.4 Task-vector construction and deployment

Construction. Rather than ascending the forget loss directly, which is unbounded and unstable, we build the edit by a well-posed descent and then reverse it. During forget-task fine-tuning, we replace the visual gradient by $\mathbf{m}_t \odot \nabla_{\boldsymbol{\theta}^{(v)}}\ell$ and keep the text branch frozen. After T epochs, we form $\mathbf{v}^{(v)} = \boldsymbol{\theta}_u^{(v)} - \boldsymbol{\theta}_0^{(v)}$ and deploy

$$\boldsymbol{\theta}^{(v)}(\alpha) = \boldsymbol{\theta}_0^{(v)} - \alpha\mathbf{v}^{(v)}, \qquad \alpha \geq 0. \tag{12}$$

Legitimacy of negation. Negation reuses a vector built for *learning* the forget task to *unlearn* it. The local analysis of Sections 3.1–3.2 shows that this reversal changes neither the selected parameters nor the retention cost.

Proposition 4 (Negation) *Let* $\mathbf{v}$ *be a task vector with* $\mathbf{g}_f^\top\mathbf{v} < 0$*, i.e., a descent direction of* ℓ_f *at* $\boldsymbol{\theta}_0$*. Then for all* $\alpha \geq 0$*:*

(i) Support invariance: $\sigma_k(\mathbf{g}_f) = \sigma_k(-\mathbf{g}_f)$ *for every* k*, so learning and unlearning share the optimal support of Proposition 2; moreover* $\mathrm{supp}(-\alpha\mathbf{v}) = \mathrm{supp}(\mathbf{v})$.

(ii) Even retain cost: $\widetilde{\mathcal{C}}_r(-\alpha\mathbf{v}) = \widetilde{\mathcal{C}}_r(\alpha\mathbf{v}) = \frac{1}{2}\alpha^2\,\mathbf{v}^\top H_r\mathbf{v}$*, and* $\mathcal{C}_r(\pm\alpha\mathbf{v}) = \frac{1}{2}\alpha^2\,\mathbf{v}^\top F_r\mathbf{v} + O(\alpha^3)$.

(iii) Odd forget effect: $\mathcal{U}_f(-\alpha\mathbf{v}) = -\mathcal{U}_f(\alpha\mathbf{v}) + O(\alpha^2) = \alpha\,|\mathbf{g}_f^\top\mathbf{v}| + O(\alpha^2) > 0$ *for sufficiently small* α.

Proposition 4 completes the local picture. The mask exploits the difference in *order* between forgetting and retention, and negation exploits their difference in *parity*. Its guarantees are local. Reversal of the forget effect beyond first order is the standard task-arithmetic assumption (Ilharco et al., 2023), which we verify empirically by checking that the forget loss increases monotonically in α. The cost $\mathbf{v}^\top H_r\mathbf{v}$ also contains cross-step terms $2\sum_{s<t}\boldsymbol{\Delta}_s^\top H_r\boldsymbol{\Delta}_t$, which per-step gating does not control (App. A.10). Finally, by Proposition 1, sweeping α moves along the curve $c(\mathbf{v})\,U^\star(\varepsilon)$. A single constructed vector therefore yields a whole forget–retain trade-off, and α selects the operating point. Figure 2 and Alg. 1 illustrate the pipeline used in our experiemnts.

Algorithm 1: TVE-Fisher pseudo-code.

1. Compute $\boldsymbol{\zeta}^{(r)}$ for $\boldsymbol{\theta}_0$ on pseudo-labels from the frozen model (Eq. (6)).
2. Initialize $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$.
3. **for** $t = 1, \ldots, T$ **do**
4. Compute $\widehat{\boldsymbol{\zeta}}_t^{(f)}$ with ground-truth forget labels at $\boldsymbol{\theta}_t$ and update $\bar{\boldsymbol{\zeta}}_t^{(f)}$ (Eqs. (6), (7)).
5. Calibrate both statistics (Eq. (8)) and compute $\mathbf{m}_t$ (Eq. (9)).
6. Fine-tune one epoch on $\mathcal{D}_f$ with gated visual gradients $\mathbf{m}_t \odot \nabla_{\boldsymbol{\theta}^{(v)}}\ell$.
7. **end for**
8. Form $\mathbf{v}^{(v)} = \boldsymbol{\theta}_u^{(v)} - \boldsymbol{\theta}_0^{(v)}$ and deploy $\boldsymbol{\theta}_0^{(v)} - \alpha\mathbf{v}^{(v)}$ (text branch frozen).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. Following the common practice, the evaluation of CLIP is conducted on eight datasets. To measure how well the edited model preserves retained visual knowledge, we use ImageNet Val. Split transductively (without labels) as the retain set. Additional dataset details are provided in Appendix B.1.

Compared methods & evaluations. We compare our method against standard task-vector unlearning Ilharco et al. (2023), task-vector relevant baselines, as well as a function-level ensemble baseline. We assess unlearning performance jointly on the forget set and the retain set: stronger unlearning is reflected by lower accuracy on the forget set, while the retain set is used to enforce the target retention level required by the protocol. More details are provided in Appendix B.2.

Experimental details. Our task vectors are derived from selectively fine-tuned CLIP models trained on each forget dataset, starting from the same pre-trained checkpoint and updating only the visual branch while keeping the text encoder fixed. To study robustness across model scales, we perform evaluation with three CLIP backbones. Additional implementation details are given in Appendix B.3.

Table 1: Main unlearning results on CLIP-based VLMs averaged over forget datasets. Lower forget accuracy indicates stronger unlearning, while retain accuracy should remain high under the fixed-retain protocol. Rank orders unlearning methods by mean displayed forget accuracy across the three backbones (lower is better); retention is reported separately.

Method	ViT-B/32		ViT-B/16		ViT-L/14		Overall F rank ↓
	Forget (↓)	Retain (↑)	Forget (↓)	Retain (↑)	Forget (↓)	Retain (↑)	
Pre-trained Model	48.35	63.34	55.75	68.34	65.58	75.54	–
Standard Task Arithmetic-FFT Ilharco et al. (2023)	24.06	60.66	20.41	64.70	16.45	72.03	11
Standard Task Arithmetic-LoRA Ilharco et al. (2023)	24.85	**60.95**	22.97	64.84	18.57	72.45	13
Standard Task Arithmetic-XGBLoRA Ilharco et al. (2023)	25.52	60.81	22.55	64.99	19.76	**72.74**	14
Best Val. Model Wortsman et al. (2022)	22.50	60.48	19.71	64.36	15.96	71.57	8
Uniform Merge Wortsman et al. (2022)	22.86	60.57	20.12	64.60	17.08	72.00	10
Greedy Merge Wortsman et al. (2022)	21.97	60.39	19.14	64.45	16.58	71.60	7
TIES-Merging Yadav et al. (2023)	23.68	60.19	19.94	64.49	17.59	72.41	12
EMR-Merging Huang et al. (2024)	21.61	60.28	18.96	64.46	15.50	71.82	5
Function Ensemble Wortsman et al. (2022); Dietterich (2000)	22.57	60.25	19.73	64.41	16.13	71.84	9
NegMerge Kim et al. (2025)	20.76	60.36	19.24	64.54	17.32	72.08	6
OrthoReg Liu et al. (2026)	18.92	60.90	17.81	**65.48**	15.93	72.54	4
Ours (F-TVE-FFT)	15.93	60.90	**15.42**	64.93	**14.46**	72.56	**1**
Ours (F-TVE-LoRA)	16.56	60.81	16.35	65.10	15.20	72.07	**3**
Ours (F-TVE-XGBLoRA)	**15.56**	60.67	15.44	65.00	14.95	71.88	**2**

Table 2: CLIP ViT-B/32 per-dataset unlearning results (standard task vectors). F denotes accuracy on the forget dataset and R denotes accuracy on the retain dataset (ImageNet). Rank orders unlearning methods by mean displayed forget accuracy across the eight datasets; retention is reported separately.

Method	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN		Overall F rank ↓
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	
Pre-trained Model	59.7	63.3	44.5	63.3	45.3	63.3	32.1	63.3	48.3	63.3	60.4	63.3	63.2	63.3	33.4	63.3	–
Standard Task Arithmetic (FFT) Ilharco et al. (2023)	34.6	60.7	29.4	60.4	11.8	60.8	8.6	60.0	16.2	60.5	31.7	61.0	51.9	60.8	8.2	61.0	12
Standard Task Arithmetic (LoRA) Ilharco et al. (2023)	38.7	60.5	30.3	59.6	11.0	61.8	8.5	**61.5**	16.4	61.8	32.6	**61.3**	52.2	59.2	9.2	**62.0**	13
Standard Task Arithmetic (XGBLoRA) Ilharco et al. (2023)	42.1	60.9	29.1	60.3	10.6	61.0	10.8	60.7	16.5	61.1	34.4	60.8	50.0	60.5	10.6	61.0	14
Best Val. Model Wortsman et al. (2022)	31.8	60.3	28.6	60.4	10.5	60.5	7.9	60.3	13.4	61.2	29.5	60.2	50.9	60.2	7.5	60.5	8
Uniform Merge Wortsman et al. (2022)	32.4	60.6	29.0	**60.7**	10.3	60.3	6.9	60.6	16.3	60.7	25.8	60.2	49.9	60.3	12.0	61.0	10
Greedy Merge Wortsman et al. (2022)	31.9	60.2	28.5	60.4	9.9	60.2	7.3	60.3	11.4	60.6	29.8	60.5	49.7	60.2	7.6	60.5	7
TIES-Merging Yadav et al. (2023)	33.3	60.2	30.2	60.2	10.4	60.0	8.1	60.2	15.5	60.5	31.2	59.9	52.3	60.3	8.1	60.2	11
EMR-Merging Huang et al. (2024)	30.8	60.0	27.6	60.3	9.6	60.1	7.3	60.3	13.7	60.4	29.0	60.3	47.7	59.8	7.3	60.7	6
Function Ensemble Wortsman et al. (2022); Dietterich (2000)	32.1	60.0	28.8	60.1	10.0	60.0	7.6	60.4	14.3	60.5	30.2	60.2	49.7	60.0	7.6	60.6	9
NegMerge Kim et al. (2025)	27.4	60.4	27.2	60.5	7.9	60.2	6.2	60.0	20.5	59.9	22.6	60.5	**47.2**	60.6	7.2	60.9	5
OrthoReg Liu et al. (2026)	37.1	**61.1**	**18.9**	59.6	**7.5**	60.7	3.8	60.9	**7.9**	61.2	20.3	61.0	47.5	**61.1**	8.4	61.5	4
Ours (F-TVE-FFT)	16.5	60.2	23.6	60.2	10.1	**62.4**	1.5	61.0	9.5	61.6	11.7	60.2	48.0	60.2	6.7	61.5	**2**
Ours (F-TVE-LoRA)	16.8	60.2	24.3	60.2	11.7	62.3	2.3	60.5	9.9	**62.3**	12.4	60.2	48.0	60.2	7.0	60.6	**3**
Ours (F-TVE-XGBLoRA)	**16.1**	60.2	24.3	60.3	9.2	62.1	**1.3**	60.4	8.2	61.2	**11.1**	60.2	48.1	60.2	**6.4**	60.7	**1**

4.2 MAIN RESULTS

4.2.1 FISHER-GUIDED IMPLEMENTATIONS.

Unlearning on VLMs. We evaluate our method in the standard parameter space and compare it against standard task arithmetic and a range of task-vector baselines across CLIP backbones. Ta-

ble 1 summarizes the average forget and retain accuracies over the eight forget datasets, where each target dataset is unlearned individually and the resulting performance is then averaged. Overall, the compared methods maintain relatively strong retain performance, indicating that the fixed retain performance is maintained by different unlearning strategies. Within this setting, our FIM-guided task-vector methods achieve the strongest unlearning result among all reported methods. In particular, on ViT-B/32, F-TVE-XGBLoRA achieves a forget accuracy of 15.56, with F-TVE-FFT and F-TVE-LoRA also getting forget accuracies of 15.93 and 16.56, respectively outperforming the strongest non-FIM baseline OrthoReg by 3.36, 2.99, and 2.36 points. This result suggests that FIM-guided selective updating of model parameters produces a more effective forgetting direction than other task-vector negation and existing vector-level merging strategies. The same trend holds for ViT-B/16 and ViT-L/14. Table 2 provides the per-dataset results for ViT-B/32; results for the other backbones are in Appendix D.

Unlearning on MLLMs.

We further apply F-TVE-LoRA to the multimodal CLEAR benchmark, which contains fictional author profiles with face image–caption pairs. Using LLaVA-1.5-7B with a CLIP ViT-L/14 vision encoder, task vectors are constructed by fine-tuning on the target image–caption pairs. Forgetting and retention are measured using identity-QA classification accuracy. F-TVE-LoRA reduces forget accuracy by 5.71 points while improving retain accuracy over TA-LoRA, extending our evaluation to multimodal generative models.

Table 3: CLEAR Identity QA (multi-modal) results.

Method	Forget (↓)	Retain (↑)
Pre-trained model	54.73	47.57
Standard Task Arithmetic-LoRA	37.25	44.59
Uniform Merge	34.97	**45.13**
TIES-Merging	33.55	44.45
EMR-Merging	32.47	43.10
Function Ensemble	33.16	44.70
MMUnlearner	32.29	43.05
Ours (F-TVE-LoRA)	**31.54**	45.10

4.3 Ablation Studies

Figure 3 shows that shared normalization, dynamic Fisher estimation, and Fisher-guided masking each improve forgetting across all reported datasets.

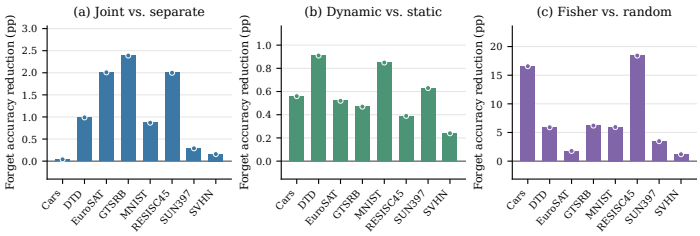


Figure 3: **Ablation gains on CLIP ViT-B/32.** Forget accuracy reduction (percentage points) relative to separate normalization (a), static Fisher (b), and random masking (c).

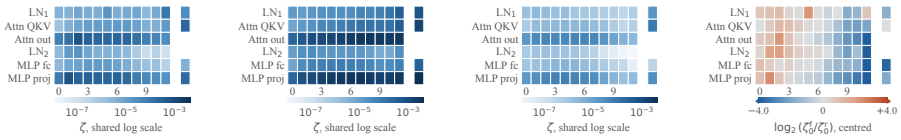
Joint vs. individual FIM normalization. Joint normalization uses a shared min–max range for forget and retain Fisher statistics, whereas separate normalization rescales each independently. The largest reduction is 2.39 points on GTSRB (Appendix E.2).

Static vs. dynamic FIM guidance. Updating forget Fisher during training outperforms fixing it at initialization, with reductions of 0.24–0.91 points across datasets (Appendix E.3).

Why FIM matters (comparison with random masking). Replacing Fisher-derived masks with random masks sampled from $[0, 1]$ weakens forgetting. Fisher guidance reduces forget accuracy by 18.43 points on RESISC45 and 16.57 points on Cars, supporting the role of parameter importance beyond masking alone (Appendix E.4).

4.4 Qualitative and Interpretability Analyses.

We visualize Fisher statistics, masks, and the unlearning trajectory to examine how forget and retain importance are distributed across the network and translated into selective parameter updates.



(a) Forget, θ_0. (b) Retain, θ_0. (c) Forget, last ep. (d) Selectivity.

Figure 4: **Layer importance profiles for CLIP ViT-B/32 on SVHN.** Each pixel is one parameter tensor, horizontal axis: transformer block, vertical axis: role inside that block, the strip on the right: the four tensors outside the residual stack.

Layer importance profiles. Figure 4 shows the layer-wise FIM statistics for the forget and retain sets. A clear pattern emerges already at initialization: the forget- and retain-wise FIM maps (Figure 4a and 4b) are not concentrated on exactly the same parts of the visual encoder (selectivity map in Figure 4d), suggesting that forgetting and retention depend on only partially overlapping parameter subspaces. This separation is precisely what makes selective unlearning possible. Comparing Figure 4a and 4c isolates that the relative ordering of the tensors is largely preserved (rank correlation 0.88), but the values themselves are not stationary, which supports dynamic FIM.

Parameter update profiles. Figure 5 visualizes the layer-wise mask profiles derived from FIM signals. The layer-wise mask maps in Figure 5 (a)–(b) show that mask-driven updates are highly structured and evolve over training, rather than being uniformly distributed across the visual encoder. The distribution view in Figure 6 (a) further shows a histogram of mask values, with an inset highlighting the rare high-value tail. The dashed q95/q99 markers indicate that the vast majority of parameters receive near-zero gates, while only a small fraction attains substantially larger update weights. This heavy-tailed, sparse gating behavior supports the intended mechanism of TVE-Fisher—forgetting is driven by a limited set of forget-salient yet retain-safe parameters, instead of broad indiscriminate editing. Finally, Figure 6 (b) shows that the task-vector negation scale α provides a smooth control: increasing α yields progressively stronger forgetting, whereas retention degrades more gradually, enabling controllable unlearning at inference time.

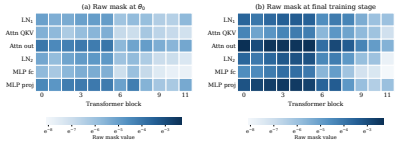
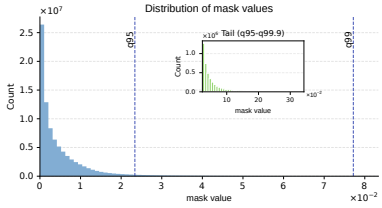
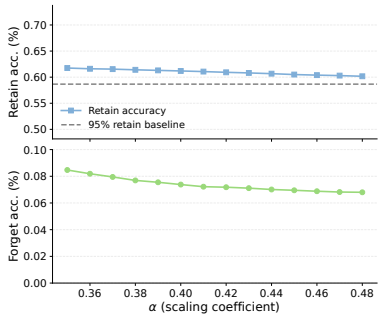


Figure 5: **Layer-wise mask profiles for CLIP ViT-B/32 on SVHN.** Module-level mean values at initialization (a) and the final training stage (b).



(a) Distribution of mask values.



(b) α-sweep of task-vector negation.

Figure 6: **Parameter update profiles and controllability.** The histogram indicates that most mask values remain small (sparse and selective gating). The α-sweep curve confirms that the strength of forgetting can be adjusted smoothly at inference time.

5 Conclusions

This paper approaches unlearning in multimodal models through the lens of FIM-based gating, inspired by the geometry of unlearning that leads to several natural mask strategies based on forget and retain FIM vectors, we select gradients that are responsive to the forgetting objective but do not interfere with must-preserve data during finetuning. Experiments show that our approach leads to a consistently better forgetting–retention trade-off compared to existing task-vector based alternatives. Our results suggest that carefully filtered task vectors by forget–retain FIM-based soft masks, can provide a lightweight, controllable, and interpretable route to precise model editing. We hope this perspective will encourage further work on information-theoretic unlearning.

Limitations & future work. One limitation of our model is due to its transductive nature (an access to a small retain set without labels is required). We ablate the model's behavior w.r.t. retain set size in Appendix F. Another trade-off is frequency of forget FIM evaluations–more frequent updates could improve accuracy but require extra computations. In the future, we will try full matrix FI approximations/layer-wise modulation strategies.

AI USE STATEMENT

No generative AI tools were used.

REPRODUCIBILITY STATEMENT

The method definitions and experimental protocol are described in Sections 3 and 4. Additional dataset configurations, baseline specifications, implementation details, per-dataset results, and ablation details are provided in the appendix to support reproduction of the reported results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Frederik Benzing. Unifying importance based regularisation methods for continual learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 2372–2396, 2022.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 141–159, 2021.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 104–120, 2020.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20089–20103, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5138–5146, 2019.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15. Springer, 2000.

Junhao Dong, Hao Zhu, Yifei Zhang, Xinghua Qu, Yew-Soon Ong, and Piotr Koniusz. Machine unlearning via task simplex arithmetic. In *Advances in Neural Information Processing Systems*, 2025.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20582–20603, 2025.

European Union. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union, L119, 2016. Article 17: Right to erasure ("right to be forgotten").

Ronald A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925. doi: 10.1017/S0305004100009580.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18695–18705, 2025.

Jiahui Geng and Qing Li. Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y. Zou. Making AI forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842, 2020.

Haochen Han, Alex Jinpeng Wang, Peijun Ye, and Fangming Liu. Unlearning the noisy correspondence makes clip more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4518–4528, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. In *NeurIPS*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Leonardo Iurada, Marco Ciccone, and Tatiana Tommasi. Efficient model editing with task-localized sparse fine-tuning. In *International Conference on Learning Representations*, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4904–4916, 2021.

Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. Approximate domain unlearning for vision-language models. In *Advances in Neural Information Processing Systems*, 2025.

Hyo Seo Kim, Dongyoon Han, and Junsuk Choe. NegMerge: Sign-consensual weight merging for machine unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30067–30085. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25f.html.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5583–5594, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwińska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013.

Alexey Kravets and Vinay Namboodiri. Zero-shot class unlearning in clip with synthetic samples. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 6456–6464, 2025.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *International Conference on Learning Representations*, 2025.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *Advances in Neural Information Processing Systems*, 2024.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Shangge Liu, Yuehan Yin, Lei Wang, Qi Fan, Yinghuan Shi, Wenbin Li, Yang Gao, and Dacheng Tao. Understanding and enforcing weight disentanglement in task arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28744–28753, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Understanding_and_Enforcing_Weight_Disentanglement_in_Task_Arithmetic_CVPR_2026_paper.html.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *International Conference on Learning Representations*, 2025.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2408–2417, 2015.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

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.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, 2018.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, Shuyang Zhu, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022.

Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4528–4537, 2018.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. doi: 10.1016/j.neunet.2012.02.016.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5385, 2024a.

Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*, 2024b. URL https://arxiv.org/abs/2405.07406.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024.

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. Cliperase: Efficient unlearning of visual-textual associations in clip. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30438–30452, 2025. doi: 10.18653/v1/2025.acl-long.1469. URL https://aclanthology.org/2025.acl-long.1469/.

Kotaro Yoshida, Yuji Naraki, Takafumi Horie, Ryosuke Yamaki, Ryotaro Shimizu, Yuki Saito, Julian McAuley, and Hiroki Naganuma. Mastering task arithmetic: τJp as a key indicator for weight disentanglement. In *International Conference on Learning Representations*, 2025.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3987–3995, 2017.

A Derivations and Proofs for Section 3

Throughout, $\mathbf{g}_f = \nabla \mathcal{U}_f(\mathbf{0})$, $H_r = F_r + \gamma I \succ 0$, $D_r = \mathrm{diag}(H_r)$, and $h_k = (H_r)_{kk}$. Statements in Sections A.1–A.7 concern the damped local surrogate; Sections A.8–A.11 concern the implementation.

A.1 Expansion of the retain cost and the damped surrogate

Write $p_0 = p_{\boldsymbol{\theta}_0}$. At $\boldsymbol{\delta} = \mathbf{0}$ the two distributions coincide, so $\mathcal{C}_r(\mathbf{0}) = 0$. The gradient vanishes because

$$\nabla \mathcal{C}_r(\mathbf{0}) = -\mathbb{E}_{\mathbf{x}} \sum_y p_0(y \mid \mathbf{x}) \nabla \log p_0(y \mid \mathbf{x}) = -\mathbb{E}_{\mathbf{x}} \nabla \sum_y p_0(y \mid \mathbf{x}) = -\nabla 1 = \mathbf{0}.$$

Since $\sum_y \nabla^2 p_0 = \nabla^2 1 = 0$, the Hessian is

$$F_r = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{y \sim p_0} \left[\nabla \log p_0 \, \nabla \log p_0^\top \right].$$

Assuming third derivatives are bounded uniformly over retained inputs near $\boldsymbol{\theta}_0$, this gives (1). The damped surrogate satisfies

$$\widetilde{\mathcal{C}}_r(\boldsymbol{\delta}) = \tfrac{1}{2} \boldsymbol{\delta}^\top F_r \boldsymbol{\delta} + \tfrac{\gamma}{2} \|\boldsymbol{\delta}\|_2^2 \;\geq\; \tfrac{1}{2} \boldsymbol{\delta}^\top F_r \boldsymbol{\delta}.$$

Every edit feasible for the surrogate budget is therefore feasible for the undamped quadratic budget, so surrogate frontiers are inner (achievable) approximations of the undamped quadratic frontier. $\widetilde{\mathcal{C}}_r$ is not itself the Taylor expansion of $\mathcal{C}_r$.

A.2 Oracle, frontier, and Proposition 1

Since $\mathbf{g}_f \neq \mathbf{0}$, the constraint is active. Stationarity of $\mathbf{g}_f^\top \boldsymbol{\delta} - \tau(\frac{1}{2} \boldsymbol{\delta}^\top H_r \boldsymbol{\delta} - \varepsilon)$ gives $\mathbf{g}_f = \tau H_r \boldsymbol{\delta}$, hence $\boldsymbol{\delta} = \tau^{-1} H_r^{-1} \mathbf{g}_f$. The active constraint gives $\tau^{-1} = \sqrt{2\varepsilon / \mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f}$, which yields (3). The frontier slope is $dU^\star / d\varepsilon = \tau \propto \varepsilon^{-1/2}$. The scalar $\mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f$ measures local forget–retain compatibility: large values mean cheap forgetting.

Proof 1 (Proof of Proposition 1) *Along $\alpha \mathbf{v}$ we have $U = \alpha |\mathbf{g}_f^\top \mathbf{v}|$ and $\varepsilon = \frac{1}{2} \alpha^2 \mathbf{v}^\top H_r \mathbf{v}$. Eliminating α gives*

$$U = \sqrt{2\varepsilon} \, \frac{|\mathbf{g}_f^\top \mathbf{v}|}{\sqrt{\mathbf{v}^\top H_r \mathbf{v}}} = c(\mathbf{v}) \, U^\star(\varepsilon), \qquad c(\mathbf{v}) = \frac{|\mathbf{g}_f^\top \mathbf{v}|}{\sqrt{(\mathbf{v}^\top H_r \mathbf{v})(\mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f)}}.$$

Because $\langle \mathbf{v}, H_r^{-1} \mathbf{g}_f \rangle_{H_r} = \mathbf{v}^\top \mathbf{g}_f$ and $\|H_r^{-1} \mathbf{g}_f\|_{H_r}^2 = \mathbf{g}_f^\top H_r^{-1} \mathbf{g}_f$, $c(\mathbf{v})$ is the absolute H_r-cosine, which is at most one by Cauchy–Schwarz. For $\mathbf{v} = \mathbf{g}_f$, Kantorovich's inequality $(\mathbf{g}^\top A \mathbf{g})(\mathbf{g}^\top A^{-1} \mathbf{g}) \leq \frac{(\kappa+1)^2}{4\kappa} \|\mathbf{g}\|^4$ gives the lower bound in (4). Equality $c(\mathbf{g}_f) = 1$ holds if and only if $\mathbf{g}_f$ is an eigenvector of H_r.

A.3 Diagonal approximation

Lemma 1 (Diagonal approximation) *Let $R_r = D_r^{-1/2} H_r D_r^{-1/2} \succ 0$ and*

$$\chi_r = \max\{1 - 1/\lambda_{\max}(R_r), \; 1/\lambda_{\min}(R_r) - 1\}.$$

Then

$$\|D_r^{1/2} (D_r^{-1} \mathbf{g}_f - H_r^{-1} \mathbf{g}_f)\|_2 \leq \chi_r \, \|D_r^{-1/2} \mathbf{g}_f\|_2. \tag{13}$$

Proof 2 *From $H_r^{-1} = D_r^{-1/2} R_r^{-1} D_r^{-1/2}$ we get $D_r^{1/2}(D_r^{-1} - H_r^{-1}) \mathbf{g}_f = (I - R_r^{-1}) D_r^{-1/2} \mathbf{g}_f$. Since $I - R_r^{-1}$ is a spectral function of the symmetric matrix R_r, its norm is $\max_i |1 - 1/\lambda_i(R_r)| = \chi_r$.*

The bound requires only $H_r \succ 0$; no condition such as $\|R_r - I\|_2 < 1$ is needed. Since $\|D_r^{-1/2} \mathbf{g}_f\|_2 = \|D_r^{1/2} D_r^{-1} \mathbf{g}_f\|_2$, it is a relative error bound in the D_r-norm.

A.4 TRAINABLE SUBSPACES AND RANK-ONE FACTORS

For $\boldsymbol{\delta} = M\mathbf{c}$ we have $\mathbf{g}_f^\top \boldsymbol{\delta} = (M^\top \mathbf{g}_f)^\top \mathbf{c}$ and $\widetilde{\mathcal{C}}_r = \frac{1}{2}\mathbf{c}^\top M^\top H_r M\mathbf{c}$. The oracle becomes $(M^\top H_r M)^{-1} M^\top \mathbf{g}_f$, which is well defined when M has full column rank.

For a rank-one update $\Delta W = \mathbf{a}\mathbf{b}^\top$, the tangent map is

$$\mathrm{vec}(d\Delta W) = J \begin{bmatrix} d\mathbf{a} \\ d\mathbf{b} \end{bmatrix}, \qquad J = [\mathbf{b} \otimes I,\ I \otimes \mathbf{a}], \tag{14}$$

so the local geometry is $J^\top H_r J$. Because $(\mathbf{a}, \mathbf{b}) \mapsto (c\mathbf{a}, \mathbf{b}/c)$ leaves ΔW unchanged, J annihilates $(\mathbf{a}, -\mathbf{b})$, and $J^\top H_r J$ is singular without a gauge constraint or a pseudoinverse. A rank-one booster therefore supplies a low-dimensional orientation, while the gate allocates motion within it.

A.5 OPTIMAL SUPPORT (PROPOSITION 2)

Proof 3 *Restricted to S under the diagonal metric, the problem is (3) with $\mathbf{g}_{f,S}$ and $(D_r)_{SS}$. Hence $V(S)^2 = 2\varepsilon\, \mathbf{g}_{f,S}^\top (D_r)_{SS}^{-1} \mathbf{g}_{f,S} = 2\varepsilon \sum_{k\in S} \sigma_k$. A sum of nonnegative terms is maximized over $|S| = s$ by its s largest terms.*

The support captures a fraction $V(S)/V(\text{all}) = (\sum_{k\in S} \sigma_k / \sum_k \sigma_k)^{1/2}$ of the diagonal frontier. Because $\sigma_k(\mathbf{g}_f) = \sigma_k(-\mathbf{g}_f)$, the optimal support is invariant to the sign of the forget gradient.

A.6 RELAXATION FROM TOP-s TO A SOFT MASK

Let

$$(\mathrm{P}_0) \quad \max_{\mathbf{z}\in\{0,1\}^d} \boldsymbol{\sigma}^\top \mathbf{z} \ \text{ s.t. } \mathbf{1}^\top \mathbf{z} \le s,$$

let (P_1) be the same problem over $[0,1]^d$, and let

$$(\mathrm{P}_\beta) \quad \max_{\mathbf{z}\in[0,1]^d,\, \mathbf{1}^\top\mathbf{z}=s} \boldsymbol{\sigma}^\top \mathbf{z} - \tfrac{1}{2\beta}\|\mathbf{z}\|_2^2.$$

Theorem 1 (Tightness) *For integer s, (P_1) has the top-s indicator as an optimal solution, so $\mathrm{OPT}(\mathrm{P}_1) = \mathrm{OPT}(\mathrm{P}_0)$. Moreover, if each coordinate is included independently with probability z_k, then $\mathbb{E}|S| = \mathbf{1}^\top \mathbf{z}$ and $\mathbb{E}\, V(S)^2 = 2\varepsilon\, \boldsymbol{\sigma}^\top \mathbf{z}$.*

Proof 4 *(P_1) is a unit-weight fractional knapsack. Moving mass from a coordinate with smaller σ to one with larger σ never decreases the objective, so filling coordinates in decreasing order of σ is optimal. The second claim follows from linearity of expectation, since $V(S)^2$ is modular.*

Theorem 2 (Regularized selection) *The solution of (P_β) is $z_k = \mathrm{clip}(\beta(\sigma_k - \nu), 0, 1)$, with ν set by the budget. For $\beta = 1/\max_k \sigma_k$ and $s = \sum_k \sigma_k / \max_k \sigma_k$, we have $\nu = 0$ and $\mathbf{z} = \boldsymbol{\sigma}/\max_k \sigma_k$. In general $\boldsymbol{\sigma}^\top \mathbf{z} \ge \mathrm{OPT}(\mathrm{P}_0) - s/(2\beta)$, and $\mathbf{z} \to$ top-s as $\beta \to \infty$ whenever $\sigma_{(s)} > \sigma_{(s+1)}$.*

Proof 5 *The objective is strongly concave. The KKT conditions for the box and budget constraints give the clipped form, which is the Euclidean projection of $\beta\boldsymbol{\sigma}$ onto the capped simplex. With $\nu = 0$ and $\beta = 1/\max \sigma$, no clipping occurs. For the value gap, let $\mathbf{z}^\star$ be the top-s indicator, which is feasible and satisfies $\|\mathbf{z}^\star\|^2 \le s$. Then*

$$\boldsymbol{\sigma}^\top \mathbf{z} \ \ge\ \boldsymbol{\sigma}^\top \mathbf{z} - \tfrac{1}{2\beta}\|\mathbf{z}\|^2 \ \ge\ \boldsymbol{\sigma}^\top \mathbf{z}^\star - \tfrac{1}{2\beta}\|\mathbf{z}^\star\|^2 \ \ge\ \mathrm{OPT} - \tfrac{s}{2\beta}.$$

For large β, any $\nu \in (\sigma_{(s+1)}, \sigma_{(s)})$ clips exactly to the top-s indicator.

The proportional mask is the closed-form, unthresholded member of this family, not its smoothest member: for fixed s, smaller β pushes $\mathbf{z}$ toward uniform.

Theorem 3 (Order preservation) *Every mask of Theorem 2 is nondecreasing in σ_k. Hence, up to ties, each superlevel set $\{k : z_k \ge c\}$ is a top-$|S|$ set and is optimal for (P_0) at its own cardinality.*

Proof 6 *$\mathrm{clip}(\beta(\sigma - \nu), 0, 1)$ is nondecreasing in σ, so each superlevel set has the form $\{k : \sigma_k \ge c'\}$. Optimality follows from Proposition 2.*

A.7 Robust selection under estimation error

Theorem 4 (Robust selection) *Suppose $\sigma_k > 0$ and $0 < a \le \hat{\sigma}_k/\sigma_k \le b < \infty$ for all k, and let $\hat{S}$ be the top-s set of $\hat{\boldsymbol{\sigma}}$. Then*

$$\sum_{\hat{S}} \sigma_k \ge \tfrac{a}{b} \sum_{S^\star} \sigma_k \qquad \text{and} \qquad V(\hat{S})/V(S^\star) \ge \sqrt{a/b}.$$

By Theorem 3, the same bound holds for every superlevel set of the corresponding soft mask.

Proof 7 *Since $\hat{S}$ maximizes $\sum \hat{\sigma}$ over sets of size s,*

$$\sum_{\hat{S}} \sigma \ge \tfrac{1}{b} \sum_{\hat{S}} \hat{\sigma} \ge \tfrac{1}{b} \sum_{S^\star} \hat{\sigma} \ge \tfrac{a}{b} \sum_{S^\star} \sigma.$$

The bound on values follows from Proposition 2.

The main sources of distortion are the following.

- *Tempering.* $\hat{\sigma}_k/\sigma_k = h_k^{1-\lambda}$, so $b/a = \kappa_h^{|1-\lambda|}$, where κ_h is the dynamic range of the denominator.
- *Gradient noise.* $\mathbb{E}[\hat{g}_k^2] = g_k^2(1 + 1/\mathrm{SNR}_k)$, so $b/a \le 1 + 1/\min_k \mathrm{SNR}_k$. Coordinates with $g_{f,k} = 0$ violate positivity, and this is the weak point of the estimator. The product of two independent half-batch gradients is unbiased for $g_{f,k}^2$. It can be negative in finite samples, which is harmless for ranking but requires truncation at zero for a soft gate.
- *Retain estimate.* The distortion is the range of $h_k/\hat{h}_k$. With labels sampled from $p_{\boldsymbol{\theta}_0}$, $\zeta_k^{(r)}$ is unbiased for $(F_r)_{kk}$; with hard pseudo-labels it is an empirical-Fisher proxy.

A.8 Calibration in tempered-ratio form

With $R = z_{\max} - z_{\min}$, expanding $\mathrm{cal}(\zeta_k^{(f)})/(1 + \mathrm{cal}(\zeta_k^{(r)}))^\lambda$ gives

$$m_k = (R+\epsilon)^{\lambda-1} \frac{\zeta_k^{(f)} - z_{\min}}{(\zeta_k^{(r)} + \gamma_t)^\lambda}, \qquad \gamma_t = z_{\max} - 2z_{\min} + \epsilon.$$

The prefactor is global and leaves rankings unchanged. The numerator shift is a multiplicative perturbation covered by Theorem 4, with $a \ge 1 - \eta$ whenever $z_{\min} \le \eta\, \zeta_k^{(f)}$ for all competing coordinates. The shift γ_t can be negative, when $z_{\min} > (z_{\max} + \epsilon)/2$, so it is not a damping parameter. The effective denominator is nevertheless bounded: since $\zeta_k^{(r)} \in [z_{\min}, z_{\max}]$,

$$\zeta_k^{(r)} + \gamma_t \in [R + \epsilon,\, 2R + \epsilon], \qquad \kappa_h \le \frac{2R+\epsilon}{R+\epsilon} < 2.$$

Consequently, the retain term reweights scores by a factor below 2^λ, and tempering costs at most a factor $2^{-|1-\lambda|/2}$ in frontier value. Calibration therefore caps retain modulation, and λ is the lever that restores it.

A.9 AdamW and alternative masks

Under an idealized stationary second-moment model that ignores momentum and moment history, a pre-normalization gate cancels: $m_k g_k / \sqrt{\mathbb{E}[(m_k g_k)^2]} = g_k / \sqrt{\mathbb{E}[g_k^2]}$. The gate then acts through selection. Coordinates with $m_k\sqrt{v_k} \gg \epsilon$ move at the full Adam scale, and coordinates with $m_k\sqrt{v_k} \ll \epsilon$ are effectively frozen. This implicit threshold is a level set of $m_k\sqrt{v_k}$ rather than of m_k, so Theorem 3 applies exactly only to explicit thresholding. Decoupled weight decay continues to move gated-out coordinates unless they are excluded from it.

Mask (i). Under SGD with gate m_k, the diagonal retain exposure is $m_k^2 \zeta_k^{(f)} \zeta_k^{(r)}$. With this penalty alone, the forget factor cancels against the benefit $\zeta_k^{(f)} m_k$. Adding an isotropic gate-space regularizer,

$$\max_{m_k} \ \zeta_k^{(f)} m_k - \tfrac{1}{2}\big(1 + \lambda \zeta_k^{(f)} \zeta_k^{(r)}\big) m_k^2,$$

gives (10).

Mask (ii). Linearizing $(1+z)^{-\lambda} \approx 1 - \lambda z$ and clipping at zero gives (11). This is accurate only for small $\lambda \zeta_k^{(r)}$; unlike Mask (iii), it zeros every coordinate with $\zeta_k^{(r)} \ge 1/\lambda$.

A.10 NEGATION: ORDER AND PARITY

The surrogate cost is exactly even, $\widetilde{\mathcal{C}}_r(\pm\alpha\mathbf{v}) = \frac{1}{2}\alpha^2\mathbf{v}^\top H_r\mathbf{v}$, while the true cost is even only to second order, $\mathcal{C}_r(\pm\alpha\mathbf{v}) = \frac{1}{2}\alpha^2\mathbf{v}^\top F_r\mathbf{v} + O(\alpha^3)$. The forget effect is odd to first order:

$$\ell_f(\boldsymbol{\theta}_0 - \alpha\mathbf{v}) = \ell_f(\boldsymbol{\theta}_0) - \alpha\nabla\ell_f(\boldsymbol{\theta}_0)^\top\mathbf{v} + O(\alpha^2).$$

It is reversed when $\nabla\ell_f(\boldsymbol{\theta}_0)^\top\mathbf{v} < 0$, which can be checked by verifying that the forget loss increases monotonically in α. Support invariance is exact at $\boldsymbol{\theta}_0$ and holds only approximately along the trajectory, since $\zeta^{(f)}$ is evaluated at $\boldsymbol{\theta}_t$. Finally, with $\mathbf{v} = \sum_t \boldsymbol{\Delta}_t$,

$$\mathbf{v}^\top H_r\mathbf{v} = \sum_t \boldsymbol{\Delta}_t^\top H_r\boldsymbol{\Delta}_t + 2\sum_{s<t} \boldsymbol{\Delta}_s^\top H_r\boldsymbol{\Delta}_t,$$

and the cross-step terms are not controlled by per-step gating.

A.11 REDUNDANCY UNDER PROPORTIONAL STATISTICS

If $\zeta_k^{(f)} = c\,\zeta_k^{(r)}$ in calibrated units, then Mask (iii) reduces to $m_k = g(z_k)$ with $z_k = \zeta_k^{(f)}$ and $g(z) = z(1 + z/c)^{-\lambda}$. Its derivative is

$$g'(z) = (1 + z/c)^{-\lambda-1}\big[1 + (1-\lambda)z/c\big].$$

For $\lambda \le 1$, g is increasing and the ranking is unchanged. For $\lambda > 1$, g peaks at $z^\star = c/(\lambda - 1)$, so on $[0, 1]$ the ranking can be reversed only if $c < \lambda - 1$. Proportionality refers to calibrated units, because the shared affine shift does not preserve raw proportionality.

B ADDITIONAL EXPERIMENTAL CONFIGURATIONS

This supplementary document provides additional details of our experimental setup for VLM and MLLM unlearning, including dataset configurations (Appendix B.1), clarifications on compared approaches (Appendix B.2), and implementation details (Appendix B.3). Our protocols follow established task-vector unlearning practice for CLIP-style models Radford et al. (2021); Ilharco et al. (2023); Dong et al. (2025).

B.1 DATASET SETUPS

Our forget suite spans a diverse set of recognition scenarios to stress-test targeted forgetting under heterogeneous visual semantics and domain shifts. It includes fine-grained categorization (Stanford Cars Krause et al. (2013)), texture recognition (DTD Cimpoi et al. (2014)), remote sensing imagery (EuroSAT Helber et al. (2019) and RESISC45 Cheng et al. (2017)), traffic sign and digit recognition (GTSRB Stallkamp et al. (2012), MNIST LeCun et al. (1998), and SVHN Netzer et al. (2011)), and scene recognition (SUN397 Xiao et al. (2010)). This dataset suite follows the common protocol used in recent CLIP/VLM unlearning studies for measuring forgetting across multiple visual domains Dong et al. (2025); Ma et al. (2025).

To quantify retained general-purpose visual capability, the retain set is fixed to the ImageNet validation split Deng et al. (2009). We adopt a transductive retain protocol: retain images are observable during unlearning, but their ground-truth labels are not assumed available. Retention is consistently reported on the same retain test split, providing a unified yardstick for comparing edits across CLIP backbones and forget datasets Radford et al. (2021); Dong et al. (2025).

Each forget dataset $\mathcal{D}_f$ defines an independent unlearning instance with a shared retain set $\mathcal{D}_r$ (ImageNet val). For every instance, Fisher statistics are estimated on $\mathcal{D}_f$ to guide selective fine-tuning and produce a forget-specific edited model, from which a visual task vector is extracted. This yields eight TVE-Fisher instances in total (one per forget dataset), while the retain evaluation protocol remains unchanged across all instances Ilharco et al. (2023); Dong et al. (2025).

Retain-side Fisher is computed once at the pre-trained initialization θ_0 using transductive pseudo labels inferred by the frozen pre-trained CLIP model, which provides a label-free surrogate for identifying retain-critical parameter sensitivity. In contrast, forget-side Fisher is recomputed once per epoch under the current parameters θ_t to track the evolving reliance on forget data as optimization proceeds. Following the fixed-retain protocol, the mask strength λ in Eq. (9) and the task-vector

negation scale α in Eq. (12) are selected on validation data; final accuracies are reported on the test sets. We report the pre-trained baseline θ_0, the selectively fine-tuned model θ_u, and the final unlearned model obtained by task-vector negation $\theta(\alpha) = \theta_0 - \alpha v$ on the visual branch Ilharco et al. (2023); Ortiz-Jimenez et al. (2023); Dong et al. (2025).

B.2 Compared methods and evaluation protocol

This section details the VLM baselines in Table 1 and their evaluation under task arithmetic. The CLEAR protocol is specified separately in Appendix B.4.

Notation. Let θ_0 denote the pre-trained CLIP weights and $\mathcal{D}_f$ a forget dataset. Selective fine-tuning on $\mathcal{D}_f$ produces θ_u, from which we extract a visual task vector $v = \theta_u^{(v)} - \theta_0^{(v)}$. Unlearning by task-vector negation applies $\theta^{(v)}(\alpha) = \theta_0^{(v)} - \alpha v$ while keeping the text branch unchanged Ilharco et al. (2023). For each forget dataset, α is chosen on a small held-out split to satisfy the fixed-retain protocol (retain accuracy meets the target), and forgetting is compared under this constraint.

Standard task arithmetic. This baseline performs unlearning by negating a single task vector obtained from fine-tuning on the forget dataset, i.e., $\theta(\alpha) = \theta_0 - \alpha v$ on the visual branch Ilharco et al. (2023). It is lightweight and retraining-free at inference time but can be sensitive to the particular fine-tuning configuration.

Vector-level merging baselines. We reinterpret representative model combination strategies as *task-vector merging* rules by treating each task vector as a parameter delta and merging deltas before negation. This follows the common view that ensembling/merging can be applied to fine-tuned checkpoints or their parameter differences Wortsman et al. (2022); Yadav et al. (2023); Huang et al. (2024). Concretely, given a set of candidate task vectors $\{v_i\}_{i=1}^{Q}$:

- **Uniform merge:** average the task vectors $\bar{v} = \frac{1}{Q}\sum_i v_i$ and unlearn via $\theta_0 - \alpha\bar{v}$ Wortsman et al. (2022).
- **Greedy merge:** iteratively add task vectors that most improve the validation unlearning score under the retain constraint, producing a merged vector used for negation Wortsman et al. (2022).
- **TIES-Merging:** trim low-magnitude entries and resolve sign disagreements across vectors before merging, mitigating destructive interference Yadav et al. (2023).
- **EMR-Merging:** elects a unified task vector and constructs task-specific masks and rescaling factors directly from the task vectors Huang et al. (2024).

NegMerge. NegMerge Kim et al. (2025) averages candidate task vectors from different fine-tuning configurations, retaining only coordinates with consistent signs across candidates. The resulting consensus vector is negated under the same retain-constrained selection protocol.

OrthoReg. OrthoReg Liu et al. (2026) regularizes weight-update matrices during fine-tuning to encourage internal orthogonality and improve task-vector disentanglement. We evaluate task negation using the resulting vectors under the same retain-constrained protocol.

Function-level ensemble baseline. To isolate the benefit of aggregating predictions rather than parameters, we consider a vanilla ensemble that averages the logits (or probabilities) of Q independently unlearned models at inference time Dietterich (2000). Each member is obtained by negating its own task vector, and predictions are averaged across members. This baseline can improve stability but increases inference cost linearly in Q.

B.3 Experimental details

All experiments are conducted with OpenAI CLIP checkpoints Radford et al. (2021) and focus on unlearning through subtraction-based task-vector editing Ilharco et al. (2023). Unless stated otherwise, the text encoder is kept frozen to preserve the pre-trained textual embedding space used for zero-shot classification, and only the visual encoder is updated to construct the forget-specific

task direction. We evaluate across three widely used CLIP variants (ViT-B/32, ViT-B/16, and ViT-L/14). Within each backbone, all compared methods initialized from the same pre-trained weights for fair comparison.

Task vectors are obtained by selectively fine-tuning on each forget dataset using the standard CLIP-style cross-entropy objective on image–text logits. Optimization uses AdamW with learning rate 1×10^{-5} and weight decay 0.1. The number of fine-tuning epochs is set per dataset and kept fixed across compared methods within each fine-tuning parameterization: 35 epochs for Cars, 76 for DTD, 12 for EuroSAT, 11 for GTSRB, 5 for MNIST, 15 for RESISC45, 14 for SUN397, and 4 for SVHN. This ensures that performance differences reflect the unlearning mechanism rather than unequal training configurations.

The retain set is fixed as the ImageNet validation split Deng et al. (2009) and is used transductively (retain images are observable but labels are not assumed). To compute retain-side Fisher at initialization θ_0, we therefore replace ground-truth labels with pseudo labels inferred by the frozen pre-trained CLIP model. Specifically, given CLIP logits $s_c(x)$ for class c, we form the posterior

$$q(c \mid x) = \frac{\exp(s_c(x))}{\sum_{j=1}^{C} \exp(s_j(x))},$$

and define the hard pseudo label

$$\hat{y}^{\text{hard}}(x) = \arg \max_{c \in \{1,\dots,C\}} q(c \mid x).$$

The retain-side loss is then

$$\mathcal{L}_{\text{retain-hard}}(x;\theta) = -\log p_\theta\big(\hat{y}^{\text{hard}}(x) \mid x\big),$$

which yields the diagonal retain Fisher estimate

$$F_i^{(\text{retain})}(\theta_0) = \mathbb{E}_{x\sim\mathcal{D}_r}\left[\left(\frac{\partial \mathcal{L}_{\text{retain-hard}}(x;\theta_0)}{\partial \theta_i}\right)^2\right].$$

Here, $F_i^{(\text{retain})}$ denotes the ith coordinate of the empirical diagonal retain statistic $\boldsymbol{\zeta}^{(r)}$ in the main text, and θ_i denotes a trainable visual-parameter coordinate.

For completeness, we also define a soft pseudo-label distribution via temperature-scaled CLIP logits,

$$q_\theta(c \mid x) = \frac{\exp\big(z_c(x)/T\big)}{\sum_{j=1}^{C} \exp\big(z_j(x)/T\big)}, \qquad z_c(x) = \tau\,\langle f_\theta(x),\, t_c\rangle,$$

To modulate the sharpness of the teacher distribution on the retained set, we further apply a power transform to the teacher probabilities restricted to the top-k classes.

Let $K(x) = \text{TopK}(q_{\theta_0}(\cdot \mid x), k)$ be the teacher top-k set and $\gamma > 0$ the power coefficient. We consider two variants for the teacher weights used in retain-FIM construction:

$$\tilde{q}_{\theta_0}^{(\text{norm})}(c \mid x) = \frac{q_{\theta_0}(c \mid x)^\gamma}{\sum_{j\in K(x)} q_{\theta_0}(j \mid x)^\gamma}, \quad c \in K(x), \tag{15}$$

and

$$\tilde{q}_{\theta_0}^{(\text{unnorm})}(c \mid x) = q_{\theta_0}(c \mid x)^\gamma, \quad c \in K(x). \tag{16}$$

In this appendix, T denotes the teacher temperature, τ the CLIP logit scale, and γ the teacher-weight power coefficient. These local meanings are distinct from the training horizon T and damping coefficient γ used in the main text.

Using these weights, the corresponding top-k soft pseudo retain-FIM takes the form

$$\boldsymbol{\zeta}^{(\text{retain})} = \mathbb{E}_{x\sim\mathcal{D}_r}\left[\sum_{c\in\mathcal{K}(x)} \widetilde{q}_{\theta_0}(c \mid x)\, \left(\nabla_{\theta^{(v)}} \log p_\theta(c \mid x)\right)^{\odot 2}\Big|_{\theta=\theta_0}\right]. \tag{17}$$

Here, p_θ is the model posterior without teacher-temperature scaling, and $\widetilde{q}_{\theta_0}$ is chosen according to Eq. (15) or Eq. (16). The teacher weights are held fixed when differentiating, and $\odot 2$ denotes element-wise squaring. For the normalized variant, $\gamma > 1$ sharpens the teacher distribution within

$\mathcal{K}(x)$, whereas $0 < \gamma < 1$ smooths it. For the unnormalized variant, γ changes both the relative weights and their total mass; at $\gamma = 1$, the weights equal the original teacher probabilities restricted to $\mathcal{K}(x)$. The normalized variant additionally rescales these weights to sum to one within $\mathcal{K}(x)$.

Forget-side Fisher is recomputed once per epoch at the current parameters θ_t as the model is fitted to $\mathcal{D}_f$; the resulting task vector is subsequently negated for unlearning. To reduce variance in the empirical Fisher estimates, we smooth the epoch-wise forget Fisher using an exponential moving average with a decay factor of 0.9 (denoted `FORGET_FIM_EPOCH_EMA_DECAY` in our implementation). In contrast, the retain Fisher is computed once at θ_0 and kept fixed throughout forget-side fine-tuning.

The Fisher-guided mask is then constructed using a small candidate set of mask strengths (e.g., $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0\}$), with the specific value selected per forget dataset on a held-out split. After selective fine-tuning, we extract the visual task vector $v = \theta_u^{(v)} - \theta_0^{(v)}$ and apply task-vector negation to obtain the final edited model. The negation scale α is chosen on the validation split following the fixed-retain protocol: we sweep a discrete set of candidate values (e.g., $\alpha \in \{0.0, 0.05, 0.1, \ldots, 1.0\}$), retain only those that preserve at least 95% of the pre-trained retain accuracy on ImageNet validation, and then select the one that minimizes forget accuracy among feasible candidates. All experiments are executed on NVIDIA V100 GPUs with 32GB memory.

B.4 CLEAR EXPERIMENTAL SETUP

We evaluate multimodal unlearning on CLEAR Dontsov et al. (2025) using LLaVA-1.5-7B with a CLIP ViT-L/14-336 vision encoder. All compared methods share the same reference checkpoint. Training uses image–caption pairs, while evaluation uses multimodal Identity QA to measure recognition of identities in the forget and retain sets. Identity QA accuracy is the fraction of questions for which the selected name matches the ground-truth identity. Lower forget accuracy indicates stronger forgetting, while higher retain accuracy indicates better preservation of non-target identities.

C ADDITIONAL DISCUSSION ON FISHER-GUIDED MASKING

Table 4 compares the three masking formulations under the AdamW fixed-retain selection protocol. Mask (iii) achieves the lowest mean forget accuracy of 17.54%, compared with 18.66% for Mask (i) and 18.17% for Mask (ii), corresponding to reductions of 1.12 and 0.63 percentage points, respectively. Mean retain accuracy remains closely comparable: 60.26% for Mask (iii), versus 60.23% and 60.28% for Masks (i) and (ii).

Table 4: Comparison of three Fisher-guided masking formulations on CLIP ViT-B/32. F: forget accuracy (↓); R: retain accuracy (↑). Bold indicates the best value among the three masks within each column.

Mask	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN		Avg.	
	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑	F ↓	R ↑
(i)	20.79	60.27	27.74	60.23	10.03	60.18	3.84	60.21	10.25	60.22	20.14	**60.25**	49.70	60.24	**6.80**	60.23	18.66	60.23
(ii)	19.05	**60.30**	27.38	60.19	9.82	**60.29**	3.98	60.27	**10.09**	**60.32**	18.95	60.23	49.18	60.26	6.91	60.34	18.17	**60.28**
(iii)	**18.53**	60.19	**26.12**	**60.27**	**8.70**	60.25	**3.56**	**60.28**	10.14	60.30	**17.52**	60.18	**48.92**	**60.27**	6.82	**60.36**	**17.54**	60.26

D PER-DATASET RESULTS IN THE STANDARD SPACE

Table 1 reports the average forget/retain accuracy obtained by unlearning each target dataset independently and then averaging the outcomes. The ViT-B/32 breakdown is presented in Table 2 in the main text; the remaining backbone-specific breakdowns are provided below.

Tables 5 and 6 report the per-dataset results for CLIP ViT-B/16 and ViT-L/14, respectively, complementing the ViT-B/32 results in Table 2. Across the three model scales, TVE-Fisher consistently achieves strong forgetting on individual target datasets while keeping ImageNet retain accuracy close to the fixed-retain threshold. This pattern suggests that the Fisher-guided gate does not rely on a specific dataset; instead, it provides a general mechanism for sharpening the forgetting direction. The stable retain performance further indicates that the improvement is not obtained by simply over-forgetting and sacrificing general visual capability, but by selectively updating parameters that are more forget-relevant and less retain-critical.

Overall, the per-dataset view reinforces the main conclusion from Table 1: retain–forget Fisher contrast is the key ingredient that turns task-vector unlearning from a coarse, global subtraction into a more targeted parameter edit.

Table 5: CLIP ViT-B/16 per-dataset unlearning results (standard task vectors). F denotes accuracy on the forget dataset and R denotes accuracy on the retain dataset (ImageNet).

Method	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN	
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)
Pre-trained Model	64.6	68.3	44.6	68.3	55.2	68.3	43.3	68.3	51.6	68.3	66.3	68.3	65.4	68.3	51.8	68.3
Standard Task Arithmetic (FFT) Ilharco et al. (2023)	27.6	64.6	23.9	64.2	12.2	64.9	7.4	64.4	9.2	65.2	27.1	64.6	49.4	64.7	6.2	64.7
Standard Task Arithmetic (LoRA) Ilharco et al. (2023)	30.2	64.1	27.9	63.1	14.3	65.9	9.5	**65.6**	11.4	66.2	28.6	65.1	52.6	63.0	9.4	65.9
Standard Task Arithmetic (XGBLoRA) Ilharco et al. (2023)	31.9	65.0	26.6	64.8	14.1	65.2	8.7	65.2	11.3	65.2	29.7	65.0	50.6	64.3	7.5	65.2
Best Val. Model Wortsman et al. (2022)	29.8	63.8	22.9	64.1	9.7	64.5	7.4	64.2	5.2	65.0	27.5	64.8	48.4	63.8	6.4	64.4
Uniform Merge Wortsman et al. (2022)	28.4	64.5	21.5	63.3	10.4	64.8	8.2	64.8	13.0	65.5	23.8	64.5	49.0	64.7	6.3	64.8
Greedy Merge Wortsman et al. (2022)	29.9	63.8	23.0	63.9	10.1	64.7	7.5	64.2	**4.0**	65.2	23.7	64.4	49.0	64.5	6.3	64.6
TIES-Merging Yadav et al. (2023)	28.1	64.5	25.5	64.5	8.8	64.3	6.9	64.5	13.1	64.8	26.4	64.2	44.2	64.6	6.9	64.5
EMR-Merging Huang et al. (2024)	26.7	64.4	24.2	64.5	8.4	64.0	6.6	64.6	12.5	64.8	25.1	64.0	**41.9**	64.4	6.6	64.6
Function Ensemble Wortsman et al. (2022); Dietterich (2000)	27.7	64.4	25.1	64.4	8.7	64.2	6.8	64.4	12.9	64.7	26.0	64.1	43.5	64.5	6.8	64.4
NegMerge Kim et al. (2025)	28.8	64.8	25.2	64.5	9.8	65.9	7.1	64.4	10.7	63.8	20.3	63.9	45.2	64.4	7.0	64.6
OrthoReg Liu et al. (2026)	36.3	**66.1**	**17.4**	**65.0**	**7.4**	65.9	4.5	65.4	8.4	66.2	18.5	**65.9**	44.6	**65.6**	5.4	63.9
Ours (F-TVE-FFT)	17.0	64.0	21.4	64.1	7.9	65.6	**1.6**	**65.6**	9.1	66.0	**13.6**	64.0	47.5	64.0	5.4	66.1
Ours (F-TVE-LoRA)	17.9	64.5	23.6	64.4	8.5	**66.0**	2.0	64.4	10.5	**66.5**	14.9	64.4	48.3	64.4	**5.2**	66.2
Ours (F-TVE-XGBLoRA)	**16.7**	64.6	19.5	64.4	7.7	64.4	2.3	64.4	9.6	66.2	14.0	64.4	48.5	64.4	**5.2**	**67.0**

Table 6: CLIP ViT-L/14 per-dataset unlearning results (standard task vectors). F denotes accuracy on the forget dataset and R denotes accuracy on the retain dataset (ImageNet).

Method	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN	
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)
Pre-trained Model	78.3	75.5	55.7	75.5	62.8	75.5	51.6	75.5	76.7	75.5	71.3	75.5	68.6	75.5	59.7	75.5
Standard Task Arithmetic (FFT) Ilharco et al. (2023)	16.7	70.7	21.6	70.6	9.3	71.8	7.5	71.8	3.2	74.5	16.7	71.7	50.3	71.7	6.5	73.2
Standard Task Arithmetic (LoRA) Ilharco et al. (2023)	20.8	72.3	23.1	71.3	11.2	73.2	8.1	**73.1**	5.6	73.4	21.3	**72.7**	50.7	70.4	7.9	73.3
Standard Task Arithmetic (XGBLoRA) Ilharco et al. (2023)	20.7	**72.8**	23.4	**72.5**	13.0	72.8	8.3	72.7	6.2	73.0	22.8	**72.7**	52.7	**72.5**	11.0	72.9
Best Val. Model Wortsman et al. (2022)	21.5	71.2	22.1	70.4	**5.2**	72.1	5.0	71.1	3.2	73.8	13.9	70.8	50.2	71.2	6.5	71.8
Uniform Merge Wortsman et al. (2022)	21.5	71.3	23.0	71.7	9.5	71.9	5.1	71.6	8.5	73.9	13.1	71.6	49.6	71.6	6.5	72.2
Greedy Merge Wortsman et al. (2022)	20.4	70.5	22.6	71.0	10.2	72.1	6.6	71.2	5.1	73.3	10.8	70.7	50.2	71.0	6.7	72.8
TIES-Merging Yadav et al. (2023)	22.2	72.5	23.7	72.4	9.8	72.2	5.3	72.3	8.8	73.6	13.5	71.2	51.1	**72.5**	6.7	72.4
EMR-Merging Huang et al. (2024)	19.5	71.7	20.8	71.9	8.6	71.6	4.6	71.7	7.7	71.9	11.8	71.9	44.9	71.6	5.9	72.0
Function Ensemble Wortsman et al. (2022); Dietterich (2000)	20.3	71.8	21.7	71.9	9.0	71.6	4.8	71.8	8.0	71.9	12.3	71.8	46.8	71.7	6.1	71.9
NegMerge Kim et al. (2025)	32.7	71.9	23.9	71.9	9.1	72.1	2.8	71.3	10.9	73.6	**8.8**	70.9	43.6	72.1	6.8	72.8
OrthoReg Liu et al. (2026)	34.4	72.6	**17.2**	70.5	5.5	73.8	3.8	**73.1**	6.8	74.1	10.1	72.6	**43.2**	71.3	6.5	72.3
Ours (F-TVE-FFT)	**11.1**	71.8	26.3	71.8	9.6	71.8	0.9	71.8	**2.0**	**75.2**	10.3	71.8	48.9	71.8	6.7	**74.4**
Ours (F-TVE-LoRA)	12.5	71.2	27.1	71.2	9.6	71.3	1.0	72.6	2.3	74.7	13.3	71.2	49.1	71.2	6.7	73.1
Ours (F-TVE-XGBLoRA)	11.6	70.8	26.7	71.0	10.3	**73.9**	**0.8**	70.8	2.4	74.2	11.3	70.8	50.9	70.8	**5.7**	72.7

E Validating the effectiveness of Fisher information

E.1 TVE-Fisher vs. Fisher (Naive)

We compare TVE-Fisher with Fisher (Naive), which constructs its mask solely from the forget Fisher diagonal. Both use full fine-tuning on CLIP ViT-B/32 with otherwise identical settings. As shown in Table 7, TVE-Fisher achieves lower forget accuracy on all eight datasets, reducing the average from 16.91% to 15.86%, while maintaining comparable retain accuracy (60.90% versus 60.87%). These results support the usefulness of the retain signal, consistent with the informativeness condition in Section 3.3.

Table 7: **CLIP ViT-B/32: TVE-Fisher vs. Fisher (Naive)** F: forget accuracy (↓); R: retain accuracy (↑). Bold denotes the better displayed value within each dataset and metric.

Method	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN		Avg.	
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)
Fisher (Naive)	18.43	**60.19**	24.84	**60.22**	11.07	**62.66**	1.45	60.18	12.48	**62.38**	12.33	**60.20**	47.96	**60.20**	6.71	60.91	16.91	60.87
TVE-Fisher	**16.49**	60.17	**23.56**	60.19	**10.11**	62.42	**1.11**	**60.98**	**9.49**	61.56	**11.67**	**60.20**	**47.74**	60.19	**6.66**	**61.52**	**15.86**	**60.90**

E.2 JOINT VS. SEPARATE FIM NORMALIZATION

This ablation examines the shared calibration design in Eq. (8). Joint min–max normalization applies common extreme values and range to the forget and retain Fisher diagonals, placing both statistics on a shared scale. Separate normalization instead rescales each diagonal using its own range, removing their original cross-FIM scale relationship.

We compare these alternatives using TVE-Fisher full fine-tuning on CLIP ViT-B/32. Both variants use the same training configuration, only the normalization scheme changes. As shown in Table 8, joint normalization achieves lower forget accuracy on all datasets. Joint normalization reduces average forget accuracy from 16.95% to 15.85% and increases average retain accuracy from 60.79% to 60.90%.

Table 8: **CLIP ViT-B/32: joint vs. separate FIM normalization.** F: forget accuracy (↓); R: retain accuracy (↑). Bold denotes the better displayed value within each dataset and metric.

Normalization	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN		Avg.	
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)
Separate min–max	16.53	**60.17**	24.55	60.18	12.12	62.41	3.50	60.20	10.36	61.39	13.67	**60.20**	48.03	**60.21**	6.82	**61.52**	16.95	60.79
Joint min–max	**16.49**	**60.17**	**23.56**	**60.19**	**10.11**	**62.42**	**1.11**	**60.98**	**9.49**	**61.56**	**11.67**	**60.20**	**47.74**	60.19	**6.66**	**61.52**	**15.85**	**60.90**

E.3 STATIC VS.DYNAMIC FORGET-FIM GUIDANCE

When forget-side Fisher is kept static, the mask is anchored to the initial sensitivity profile of $\mathcal{D}_f$ and cannot adapt to changes in parameter reliance along the fine-tuning trajectory. In the benefit-landscape interpretation, b_t denotes the coordinate-wise forget-benefit proxy represented by the forget-side Fisher estimate at epoch t. Static guidance holds this proxy at its initialization value b_0 throughout fine-tuning, even when the current forget-relevant directions have shifted. Dynamic recomputation instead refreshes $\widehat{\boldsymbol{\zeta}}_t^{(f)}$ once per epoch, with EMA decay, allowing the mask to track the evolving sensitivity profile while smoothing fluctuations between successive estimates.

To isolate the contribution of this refresh, we compare dynamic guidance against a static variant that computes forget Fisher only once at θ_0. Table 9 shows that dynamic guidance improves forgetting on all eight datasets on CLIP ViT-B/32, without sacrificing average retain accuracy. Average forget accuracy decreases from 16.43% to 15.86%, while average retain accuracy increases from 60.37% to 60.90%. The evidence supports the practical value of recomputing the forget-side Fisher information.

Table 9: **CLIP ViT-B/32: static vs. dynamic forget-Fisher guidance.** Static fixes the forget Fisher at initialization; dynamic recomputes it each epoch with EMA decay 0.9. Retain Fisher remains fixed in both cases. F: forget accuracy (↓); R: retain accuracy (↑), in percent. Each variant independently selects α using the same validation protocol. Bold denotes the better displayed value within each dataset and metric.

Forget-Fisher guidance	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN		Avg.	
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)
Static (θ_0)	17.05	**60.19**	24.47	**60.27**	10.63	61.28	1.58	60.44	10.34	60.20	12.06	60.19	48.37	**60.23**	6.90	60.18	16.43	60.37
Dynamic (epoch EMA)	**16.49**	60.17	**23.56**	60.19	**10.11**	**62.42**	**1.11**	**60.98**	**9.49**	**61.56**	**11.67**	**60.20**	**47.74**	60.19	**6.66**	**61.52**	**15.86**	**60.90**

E.4 WHY FIM MATTERS BEYOND MASKING

A random mask preserves dense gating but removes Fisher-informed structure. In TVE-Fisher, the update gate is shaped by where the forget objective concentrates and where the retain objective is fragile, so the mask acts as a targeted allocator rather than a generic regularizer.

Table 10 supports this distinction. FIM-guided masks achieve markedly lower forget accuracy on all eight datasets, reducing average forget accuracy from 23.30% to 15.85%, while increasing average retain accuracy from 60.53% to 60.90%.

Table 10: **CLIP ViT-B/32: FIM-guided mask vs. random mask.** Random masks are sampled independently per coordinate. FIM-guided masks use dynamic Fisher and subsequent mask values. F: forget accuracy (↓); R: retain accuracy (↑), in percent. Bold denotes the better displayed value within each dataset and metric.

Mask	Cars		DTD		EuroSAT		GTSRB		MNIST		RESISC45		SUN397		SVHN		Avg.	
	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)	F (↓)	R (↑)
Random mask	33.06	**60.81**	29.47	**60.39**	11.89	60.97	7.32	60.24	15.44	60.56	30.10	60.18	51.23	**60.85**	7.87	60.26	23.30	60.53
FIM-guided mask	**16.49**	60.17	**23.56**	60.19	**10.11**	**62.42**	**1.11**	**60.98**	**9.49**	**61.56**	**11.67**	**60.20**	**47.74**	60.19	**6.66**	**61.52**	**15.85**	**60.90**

F Limitations and additional ablations

F.1 Transductive nature of retain-side Fisher

We examine the transductive assumption in TVE-Fisher by varying the supervision used to construct the retain-side Fisher on ImageNet validation. In deployment, ImageNet labels may be unavailable, so we compare three retain-FIM variants: (i) ground-truth labels, (ii) hard pseudo labels inferred by the frozen pretrained CLIP model, and (iii) soft pseudo labels given by the frozen model's class-posterior distribution. Throughout this ablation, retain-FIM is computed on the full ImageNet-val split and stabilized by an EMA over mini-batches, while the forget-side training protocol is unchanged.

Table 11: **Effect of retain-label choices for retain-FIM construction.** Forget and retain report the average accuracy (%) over the eight forget datasets and the retain dataset (ImageNet). Best results in each column are bolded.

Method	ViT-B/32		ViT-B/16		ViT-L/14	
	Forget (↓)	Retain (↑)	Forget (↓)	Retain (↑)	Forget (↓)	Retain (↑)
TVE-Fisher_gt-labels	17.54	60.20	16.70	64.96	16.07	71.81
TVE-Fisher_hard-pseudo	**17.28**	60.22	**15.42**	64.93	**14.92**	71.84
TVE-Fisher_soft-pseudo	18.06	60.19	16.83	64.93	15.56	71.82
TVE-Fisher_soft-power-norm	19.20	**60.35**	18.15	**65.07**	17.28	71.80
TVE-Fisher_soft-power-unnorm	19.00	60.24	16.74	64.93	19.23	**71.91**

Retain loss and diagonal Fisher. For an image x, CLIP forms class logits

$$z_c^{(\theta)}(x) = s \cdot \left\langle f_\theta^{\text{img}}(x),\, t_c \right\rangle, \qquad c = 1, \ldots, C,$$

where $f_\theta^{\text{img}}(x)$ is the image feature, t_c is the class text feature, and s is the logit scale Radford et al. (2021). The hard-label variants use squared per-example gradients. The soft variant instead averages squared class-conditional scores under a fixed teacher distribution.

For the hard-label variants, the diagonal estimate is

$$F_i^{(\text{retain})} = \mathbb{E}_{x \sim \mathcal{D}_{\text{retain}}} \left[\left(\frac{\partial \mathcal{L}_{\text{retain}}(x;\theta)}{\partial \theta_i} \right)^2 \Bigg|_{\theta=\theta_0} \right].$$

Here, $\mathcal{D}_{\text{retain}} = \mathcal{D}_r$, and $F_i^{(\text{retain})}$ denotes a coordinate of the empirical diagonal retain statistic, not the full predictive Fisher F_r. In this subsection, s is the CLIP logit scale and τ is the teacher temperature, corresponding respectively to τ and T in Appendix B.3.

For the hard-label variants, we maintain an EMA over mini-batches $\mathcal{B}$ during estimation at initialization:

$$F_i^{(\text{retain})} \leftarrow \mu F_i^{(\text{retain})} + (1-\mu) \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \left(\frac{\partial \mathcal{L}_{\text{retain}}(x;\theta)}{\partial \theta_i} \right)^2 \Bigg|_{\theta=\theta_0}.$$

Here, μ is the minibatch EMA decay used during retain-statistic estimation. Retain statistics are estimated at θ_0 and then held fixed during forget-task fine-tuning.

(i) Ground-truth labels. With annotated ImageNet label $y(x)$, the retain objective is the standard cross-entropy:

$$\mathcal{L}^{\mathrm{gt}}_{\mathrm{retain}}(x) = \mathrm{CE}\Big(z^{(\theta)}(x),\, y(x)\Big)\,, \qquad F_i^{\mathrm{retain,gt}} = \mathbb{E}_x\left[\left(\frac{\partial \mathcal{L}^{\mathrm{gt}}_{\mathrm{retain}}(x)}{\partial \theta_i}\right)^2\right].$$

(ii) Hard pseudo labels. Given logits from the frozen pretrained model θ_0, the hard pseudo label is defined by

$$\hat{y}^{\mathrm{hard}}(x) = \arg\max_{c\in\{1,\dots,C\}} z_c^{(\theta_0)}(x).$$

We then replace the unknown retain label with $\hat{y}^{\mathrm{hard}}(x)$:

$$\mathcal{L}^{\mathrm{hard}}_{\mathrm{retain}}(x) = \mathrm{CE}\Big(z^{(\theta)}(x),\, \hat{y}^{\mathrm{hard}}(x)\Big)\,, \qquad F_i^{\mathrm{retain,hard}} = \mathbb{E}_x\left[\left(\frac{\partial \mathcal{L}^{\mathrm{hard}}_{\mathrm{retain}}(x)}{\partial \theta_i}\right)^2\right].$$

(iii) Soft pseudo labels. We define the teacher distribution with temperature $\tau > 0$:

$$q(c \mid x) = \frac{\exp(z_c^{(\theta_0)}(x)/\tau)}{\sum_{j=1}^{C} \exp(z_j^{(\theta_0)}(x)/\tau)}.$$

A teacher-weighted diagonal Fisher estimate is

$$F_i^{\mathrm{retain,soft}} = \mathbb{E}_x\left[\sum_{c=1}^{C} q(c \mid x)\left(\frac{\partial \log p_\theta(c \mid x)}{\partial \theta_i}\right)^2\right],$$

optionally approximated using a top-K truncation over teacher classes.

Pseudo-label quality and retain-FIM similarity (ViT-B/16). Table 12 reports pseudo-label agreement on ImageNet-val for CLIP ViT-B/16, together with similarity between the soft-pseudo retain-FIM and the GT retain-FIM. Hard pseudo labels achieve higher top-1/top-5 agreement (68.32/91.87) than soft pseudo (65.08/88.38), while the soft-pseudo retain-FIM remains highly consistent with the GT retain-FIM (Pearson/Spearman/Cosine ≈ 0.97 at the parameter level and ≈ 1.00 at the layer level). These results support pseudo supervision as a practical transductive surrogate for retain-FIM construction when retain labels are unavailable.

Table 12: Pseudo-label quality on ImageNet-val and retain-FIM similarity (CLIP ViT-B/16).

Pseudo supervision (teacher θ_0)	**Top-1 (%)**	**Top-5 (%)**
Hard pseudo labels	68.32	91.87
Soft pseudo labels	65.08	88.38

FIM similarity (Soft vs. GT)	**Pearson**	**Spearman**	**Cosine**
Parameter-level (diag. Fisher)	0.9766	0.9736	0.9766
Layer-level (aggregated)	0.9998	0.9979	0.9996

Downstream impact on unlearning (MNIST forget, $\lambda = 0.1$, ViT-B/16). We next evaluate how the retain-FIM supervision choice affects unlearning under a fixed forget instance (MNIST) and the same mask strength $\lambda = 0.1$. Following the fixed-retain protocol, α is selected on a held-out split to satisfy the retain constraint, and forgetting is compared under this constraint. Table 13 summarizes the resulting forget/retain accuracies. Compared to GT retain-FIM (forget $= 8.55$, retain $= 65.42$), hard pseudo retain-FIM yields comparable retention with slightly stronger forgetting (forget $= 7.95$, retain $= 65.03$). Soft pseudo retain-FIM remains feasible but produces weaker forgetting in this configuration (forget $= 12.57$, retain $= 64.93$). Overall, pseudo-label-based retain-FIM does not undermine retention-aware unlearning and can closely match the oracle variant in this case.

Table 13: Effect of retain-FIM supervision on unlearning (e.g. CLIP ViT-B/16, MNIST forget, $\lambda = 0.1$). α is selected by the fixed-retain protocol.

Retain-FIM supervision	**Forget Acc. (↓)**	**Retain Acc. (↑)**
GT labels	8.55	65.42
Hard pseudo	7.95	65.03
Soft pseudo	12.57	64.93

F.2 Retain set size ablation under the transductive protocol

The retain-side Fisher in TVE-Fisher is estimated transductively on unlabeled ImageNet-val by replacing ground-truth labels with pseudo labels inferred by the frozen pretrained CLIP model. This ablation varies the *amount of retain data* used to construct pseudo labels and compute retain-FIM, while evaluating retention on the held-out remainder of ImageNet-val.

We consider five split settings. A fraction $\rho \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ of ImageNet-val is used to generate pseudo labels and compute retain-FIM at initialization, and the complementary fraction $(1-\rho)$ is used exclusively for retain-accuracy evaluation. The forget task is fixed as MNIST and we use CLIP ViT-B/16 with the same unlearning configuration ($\lambda = 0.1$) across all settings. We report the best forget/retain accuracies selected under the same fixed-retain protocol.

Table 14 summarizes the results for both hard and soft pseudo-label variants. Across all split ratios, retention remains stable (around 64.8–65.1), indicating that the retain constraint is not overly sensitive to moderate changes in retain-FIM estimation set size. Forgetting shows mild variation across splits. Hard pseudo labels tend to yield slightly stronger or comparable forgetting than soft pseudo labels at similar retention, consistent with hard supervision providing a sharper retain-side importance signal under this configuration.

Table 14: Retain set size ablation for transductive retain-FIM construction (CLIP ViT-B/16; MNIST forget; $\lambda = 0.1$). A fraction ρ of ImageNet-val is used to compute retain-FIM using pseudo labels, and the remaining $(1-\rho)$ is used for retain evaluation. Forget/retain accuracies are in %. Lower forget accuracy indicates stronger unlearning.

FIM/Eval split	Hard pseudo retain-FIM		Soft pseudo retain-FIM	
	Forget (↓)	Retain (↑)	Forget (↓)	Retain (↑)
30/70	9.61	64.86	11.98	64.99
40/60	11.40	64.87	10.97	64.86
50/50	9.16	64.81	12.32	64.86
60/40	9.57	65.02	11.76	64.79
70/30	11.00	64.83	9.91	65.05

F.3 Limitations and future directions

A key limitation of TVE-Fisher is its transductive requirement: constructing the retain-side Fisher signal assumes access to a small unlabeled retain set (e.g., a subset of ImageNet-val) to estimate which parameters are retention-critical, which may be unavailable in certain deployment or data-governance settings. Additioly, while we modulate updates at the parameter level, we have not yet explored more structured alternatives such as layer-wise modulation strategies that may offer a better compute–control trade-off and improved scalability for larger Multimodal backbones.