Open-Vocabulary Domain Unlearning

Sumanth Udupa[1] **Mehrtash Harandi**[2] **Yadan Luo**[1] **Mahsa Baktashmotlagh**[1]

[1]The University of Queensland [2]Monash University

Abstract

Vision-Language Models (VLMs) exhibit remarkable zero-shot generalization, yet they often encode unwanted or hazardous stylistic domains such as idealized textbook diagrams in medical AI or cartoon vehicles in autonomous driving. Approximate Domain Unlearning (ADU) aims to selectively erase a model's recognition of a target visual domain while preserving accuracy on the remaining domains. However, existing ADU methods operate under a flawed closed-vocabulary assumption: they evaluate unlearning solely on the specific object classes seen during the unlearning fine-tuning phase. Consequently, these methods do not unlearn the domain itself; they merely overfit to seen class-domain pairs, leaving the domain easily recognizable for unseen classes and providing a false sense of removal. We argue that true domain erasure must be class-agnostic. To address this, we formalize **Open-Vocabulary Domain Unlearning (OVDU)**, a rigorous protocol that requires domain forgetting to transfer to held-out classes in a few-shot setup. To solve the OVDU challenge, we propose a surgical parameter-editing framework. First, a Fisher Information mask restricts updates to parameters with greater empirical sensitivity to the forget domain than to the retain domains, helping limit interference with zero-shot generalization. Second, our **Targeted Manifold Scattering (TMS)** objective uses preference-based mining to locally scatter the forget domain's stylistic geometry. Evaluated across PACS, OfficeHome, and DomainNet, our method vastly improves open-vocabulary generalization over existing baselines. Crucially, it delivers exceptional sample efficiency, outperforming peak 8-shot baseline results with only 4 shots. Project page: `https://sumanth181099.github.io/ovdu` .

1 Introduction

Pre-trained Vision-Language Models (VLMs) Radford et al. [2021], Faghri et al. [2025] exhibit remarkable zero-shot generalization, enabling them to recognize a vast array of open-vocabulary concepts without additional training. However, this strong generalization also means VLMs encode and retain information that may be unnecessary, computationally wasteful, or hazardous in specific downstream applications Wang et al. [2021], Shokri et al. [2017]. To address this, approximate machine unlearning has emerged to selectively remove specific knowledge while preserving overall model utility Bourtoule et al. [2021], Graves et al. [2021]. Historically, approximate unlearning in VLMs has predominantly focused on *class unlearning* wherein the task is to degrade recognition accuracy for specific object categories while maintaining accuracy for others Golatkar et al. [2020], Fan et al. [2023], Huang et al. [2024], Kuwana et al. [2024].

However, simply erasing an object class is often insufficient for real-world deployments when using pretrained VLMs. Consider an autonomous driving system: the model must reliably detect real physical cars to prevent collisions, but it must strictly avoid recognizing illustrated cars depicted on roadside advertisements as real vehicles Kawamura et al. [2025]. Standard class unlearning would blindly erase the concept of "car" entirely. This necessity has motivated the recent introduction of Approximate Domain Unlearning (ADU) Kawamura et al. [2025], which aims to selectively reduce

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

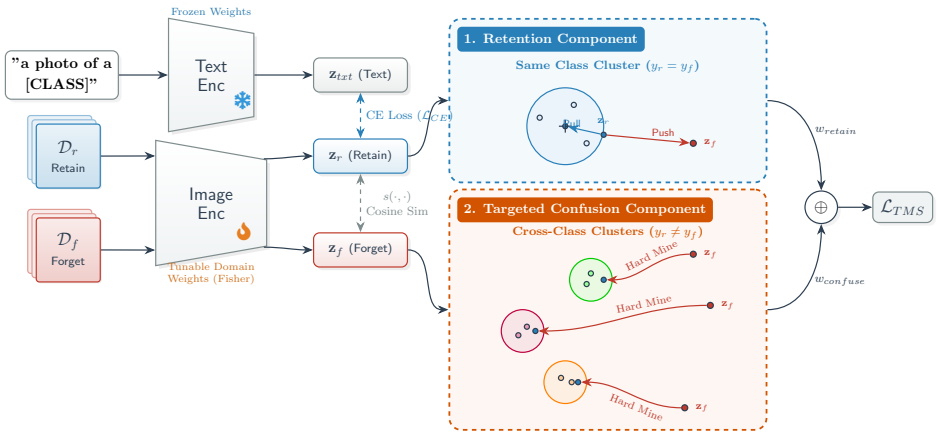

Figure 1: **Overview of the Targeted Manifold Scattering (TMS) framework.** Images from the retain ($\mathcal{D}_r$) and forget ($\mathcal{D}_f$) domains are processed through a Fisher-masked Image Encoder to isolate domain-specific updates. The TMS loss applies two decoupled geometric forces: (1) The **Retention Component** protects semantic integrity by contrasting same-class retain and forget embeddings ($y_r = y_f$). (2) The **Targeted Confusion Component** actively diffuses the forget representations by hard-mining cross-class retain clusters ($y_r \neq y_f$) and scattering the forget embeddings (z_f) into these unrelated regions, achieving efficient erasure without global semantic collapse.

recognition accuracy for images from specified "forget" domains while preserving accuracy for "retain" domains.

This requirement extends to high-stakes fields such as medical AI. Clinical VLMs pre-trained on internet-scale data often ingest "textbook diagrams" alongside real radiological scans Tu et al. [2024]. Because diagrams contain artificial shortcuts such as idealized geometries, color-coding, and bright arrows, models can learn spurious correlations Geirhos et al. [2020], DeGrave et al. [2021], leading to hallucinated diagnoses when deployed on noisy, real-world patient X-rays. To ensure safety, the "diagram" domain must be entirely unlearned.

Despite the clear practical motivation for domain unlearning, we identify a fundamental flaw in its current formulation and evaluation. Existing ADU methodologies typically evaluate unlearning performance on the exact same set of object classes that were presented during the unlearning fine-tuning phase Kawamura et al. [2025]. This closed-vocabulary protocol masks a critical failure mode: real-world environments are inherently open-ended. If an autonomous system is fine-tuned to unlearn illustrations using only common traffic classes, it will fail to generalize that domain erasure when encountering a billboard featuring a novel concept, such as a cartoon dinosaur. Similarly, in the medical domain, it is practically impossible to provide an unlearning algorithm with diagrammatic examples of every known rare disease during fine-tuning. If a domain erasure algorithm entangles stylistic features with the specific semantic classes seen during training, the VLM loses its foundational zero-shot ability to decouple pedagogical style from medical semantics on unseen, rare conditions. Furthermore, we demonstrate mathematically that current geometric regularizers exacerbate this catastrophic forgetting. Methods relying on maximizing the Maximum Mean Discrepancy (MMD) Gretton et al. [2012] to explicitly separate domain distributions in the latent space inadvertently compute a global displacement vector heavily biased by the semantic centroids of the seen unlearning classes. When applied via prompt tuning Kawamura et al. [2025], Zhou et al. [2022], MMD applies a global affine shift that misaligns and destroys the zero-shot generalization manifold for unseen classes. Similarly, rigid geometric regularizers like the hard-margin Triplet loss Hoffer and Ailon [2015] apply step-function gradients that treat all violations equally once a threshold is crossed. This induces gradient oscillation and promotes *manifold translation* wherein the forget domain is relocated into an isolated space rather than truly erasing its stylistic structure via scattering the representations locally.

To overcome these fundamental limitations, we formalize **Open-Vocabulary Domain Unlearning (OVDU)**. We argue that true domain unlearning must be class-agnostic, enabling a model to unlearn

a domain's stylistic properties using only a small proxy subset of classes ($\mathcal{C}_{train}$) and successfully generalizing that erasure to entirely unseen concepts ($\mathcal{C}_{test}$). To achieve this, we build upon Fisher-guided unlearning principles Golatkar et al. [2020], Liu et al. [2023] to restrict updates to parameters that are disproportionately sensitive to the forget domain. This provides a computationally tractable local selectivity prior, with preservation of unseen-class recognition assessed empirically. To achieve true domain erasure, we introduce **Targeted Manifold Scattering (TMS)**. Unlike MMD or hard hinge based Triplet Loss, TMS utilizes a graduated log-sigmoid formulation coupled with hard mining to force forget embeddings to actively mimic unrelated cross-class retain embeddings. This induces bounded, localized *manifold diffusion*, effectively shattering the stylistic geometry of the forget domain without disturbing global semantic alignment.

Our main contributions are summarized as follows:

• **A Novel Problem Setting and Protocol:** We identify the data leakage in current domain unlearning literature and introduce the Open-Vocabulary Domain Unlearning (OVDU) protocol, which rigorously evaluates zero-shot generalization by separating unlearning classes ($\mathcal{C}_{train}$) from testing classes ($\mathcal{C}_{test}$).
• **Targeted Manifold Scattering via Parameter Isolation:** As our primary contribution, we introduce a continuous geometric objective, Targeted Manifold Scattering, which decouples retention and targeted confusion forces to actively scatter the stylistic geometry of the forget domain. Additionally, we restrict gradient flow through a Fisher Information mask which biases optimization toward parameters with greater forget-domain than retain-domain sensitivity. Our experiments demonstrate improved forgetting–retention trade-offs, including on classes absent during unlearning.
• **Few-shot Sample Efficiency:** Extensive evaluations across the three different domain generalization benchmarks validates that our framework decisively outperforms existing baselines under the rigorous OVDU protocol. Beyond superior zero-shot domain erasure on held-out classes, our approach yields exceptional sample efficiency, nearly halving the data requirements of prior methods.

2 Related Work

Machine Unlearning. Parameter isolation and weight editing are widely used for selective forgetting, including Task Arithmetic Ilharco et al. [2022], NegMerge Kim et al. [2025], and Fisher masking Liu et al. [2023]. Related parameter-importance methods in continual learning, notably Elastic Weight Consolidation (EWC) Kirkpatrick et al. [2017], use the diagonal Fisher to identify parameters important to previously learned tasks and discourage their modification through a soft quadratic penalty. Our approach instead constructs a hard update mask from a normalized two-population ratio of forget- and retain-domain Fisher sensitivities, selecting parameters disproportionately sensitive to the forget domain. This relative sensitivity is particularly relevant to domain unlearning, where stylistic and semantic information are highly entangled (as illustrated in Figure 2).

Approximate Domain Unlearning (ADU). Recent ADU frameworks Kawamura et al. [2025] address domain erasure using prompt tuning and Maximum Mean Discrepancy (MMD) Gretton et al. [2012] for global feature separation. Yet, these methods rely on a closed-vocabulary assumption. In open-vocabulary scenarios, prompt-based MMD induces a severe global affine shift that degrades zero-shot generalization. Conversely, our approach abandons prompt tuning, leveraging Fisher-guided surgical parameter editing to isolate and degrade stylistic weights without disrupting foundational semantics.

Geometric Regularization. Rigid metric learning objectives (e.g., Triplet Loss Hoffer and Ailon [2015]) typically cause simple *manifold translation* rather than true feature disentanglement. Inspired by Direct Preference Optimization (DPO) Rafailov et al. [2023] and Negative Preference Optimization (NPO) Zhang et al. [2024], we introduce Targeted Manifold Scattering (TMS). TMS transitions preference optimization to continuous geometric editing, using hard mining to locally shatter the forget domain's stylistic geometry while seeking to retain cross-class generalization.

For a comprehensive discussion of related literature, extended theoretical comparisons, and unlearning mechanics, please refer to Appendix A.1.

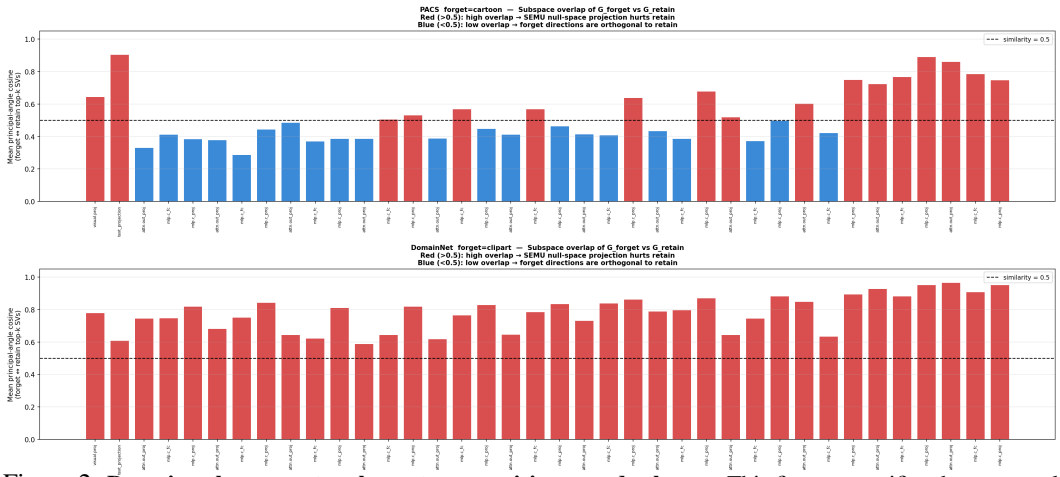

Figure 2: **Domain subspace entanglement across vision encoder layers.** This figure quantifies the structural overlap between domains in pre-trained VLMs. The x-axis represents the progressive layers of the vision encoder, while the y-axis measures the mean principal angle cosine, which specifically captures the overlap between the top k-singular vectors of the extracted domain subspaces (Forget and Retain). The prominent red bars denote exceptionally high structural similarity between the targeted forget domain and the retain domains across nearly all layers.

3 Methodology

3.1 Problem Formulation: Open-Vocabulary Domain Unlearning

Let dataset $\mathcal{D} = \{(x_i, y_i, d_i)\}_{i=1}^{N}$ comprise inputs x, class labels $y \in \mathcal{C}$, and domain labels $d \in \{\mathcal{D}_f, \mathcal{D}_r\}$, where $\mathcal{D}_f$ and $\mathcal{D}_r$ denote the forget and retain domains, respectively. Given a pre-trained VLM parameterized by θ, domain unlearning yields $\theta_{\text{unlearned}}$ that severely degrades recognition on $\mathcal{D}_f$ while strictly preserving $\mathcal{D}_r$. Standard Approximate Domain Unlearning (ADU) Kawamura et al. [2025] assumes a closed-vocabulary setting where the class availability fraction exposed during fine-tuning is $\gamma = |\mathcal{C}_{\text{train}}|/|\mathcal{C}| = 1.0$. To evaluate true stylistic decoupling, we introduce the **Open-Vocabulary Domain Unlearning (OVDU)** protocol, partitioning $\mathcal{C}$ into mutually exclusive seen classes ($\mathcal{C}_{\text{seen}}$, used strictly for training) and held-out unseen classes ($\mathcal{C}_{\text{unseen}}$).

To systematically test open-vocabulary robustness, we vary $\gamma = |\mathcal{C}_{\text{seen}}|/|\mathcal{C}| \in \{0.25, 0.50, 0.75, 1.0\}$. Crucially, OVDU evaluates on the global space $\mathcal{C}_{\text{test}} = \mathcal{C}_{\text{seen}} \cup \mathcal{C}_{\text{unseen}} = \mathcal{C}$ to verify that domain erasure is class-agnostic. For any evaluation sample $(x_{\text{test}}, y_{\text{test}}) \sim \mathcal{C}_{\text{test}}$, the model must simultaneously achieve **(1) Global Retention:** strictly preserving structural alignment and accuracy for $x_{\text{test}} \in \mathcal{D}_r$ regardless of whether y_{test} is seen or unseen, and **(2) Global Erasure:** uniformly failing to recognize the stylistic rendering of $x_{\text{test}} \in \mathcal{D}_f$ across the entire test space $\mathcal{C}_{\text{test}}$. Achieving this decoupling requires limiting interference with representations shared across domains and $\mathcal{C}_{\text{seen}}$. We therefore introduce a parameter-isolation baseline governed by empirical Fisher Information, which favors updates to parameters with greater forget-domain than retain-domain sensitivity.

3.2 Semantic Preservation via Fisher-Guided Parameter Isolation

To rigorously evaluate the necessity of surgical weight editing over global prompt tuning, we first formalize a foundational parameter-isolation baseline: Gradient Ascent (NegGrad) Graves et al. [2021] governed by an empirical Fisher Information Mask Liu et al. [2023]. An unconstrained NegGrad objective aims to induce forgetting by maximizing the cross-entropy loss on the forget domain $\mathcal{D}_f$ while minimizing it on the retain domain $\mathcal{D}_r$:

$$g_i^{\text{NegGrad}} = \nabla_{\theta_i} \mathcal{L}_{\text{CE}}(x_r, y_r; \theta) - \lambda \nabla_{\theta_i} \mathcal{L}_{\text{CE}}(x_f, y_f; \theta) \tag{1}$$

where $(x_r, y_r) \in \mathcal{D}_r$, $(x_f, y_f) \in \mathcal{D}_f$, and λ is a scaling hyperparameter. While this unconstrained gradient difference effectively erases domain recognition, it indiscriminately penalizes shared semantic weights, destroying the pre-trained VLM's zero-shot generalization capabilities on unseen classes of the retain domains Kumar et al. [2022].

To reduce this semantic collapse, parameter isolation is employed to decouple the weights responsible predominantly for stylistic domain rendering from those encoding generalized semantics. Let $\ell_j(\theta)$ denote the per-sample loss for a given instance $(x_j, y_j) \in \mathcal{D}$. We compute the diagonal empirical Fisher Information Matrix by aggregating the squared gradients of this per-sample loss over the N samples in the dataset:

$$F_{\mathcal{D}}(i) = \frac{1}{N} \sum_{j=1}^{N} \left(\nabla_{\theta_i} \ell_j(\theta)\right)^2 \tag{2}$$

Crucially, we utilize the average of the squared gradients across the batch samples rather than the maximum. This design choice ensures that the resulting sensitivity scores capture a stable, domain-wide parameter importance distribution without being disproportionately skewed by individual outlier samples or dominant object classes. We then quantify domain sensitivity for each parameter θ_i using the ratio:

$$\rho_i = \frac{F_{\mathcal{D}_f}(i) - F_{\mathcal{D}_r}(i)}{F_{\mathcal{D}_f}(i) + F_{\mathcal{D}_r}(i) + \epsilon} \tag{3}$$

where ϵ is set to a small value to ensure numerical stability.

This normalization elegantly handles domain unlearning because the forget dataset $\mathcal{D}_f$ and retain dataset $\mathcal{D}_r$ share the underlying semantic classes ($\mathcal{C}_{\text{train}}$). Parameters encoding fundamental semantics remain active in both domains, yielding $\rho_i \leq 0$, while parameters mostly dictating stylistic artifacts dominant in the forget domain, drive $\rho_i \to 1$. We isolate these predominantly forget domain sensitive stylistic weights via a binary mask with a predefined sparsity threshold τ:

$$m_i = \mathbb{1}[\rho_i > \tau] \in \{0, 1\} \tag{4}$$

Routing NegGrad updates strictly through this mask ($\theta_i \leftarrow \theta_i - \eta(g_i^{\text{NegGrad}} \cdot m_i)$) restricts optimization to coordinates with greater empirical forget-domain than retain-domain sensitivity. Specifically, for each selected parameter, Eq. (4) with $\tau = 0.3$ implies

$$\frac{F_{\mathcal{D}_f}(i) - F_{\mathcal{D}_r}(i)}{F_{\mathcal{D}_f}(i) + F_{\mathcal{D}_r}(i) + \epsilon} > 0.3 \quad \Longrightarrow \quad F_{\mathcal{D}_r}(i) < \frac{7}{13} F_{\mathcal{D}_f}(i) - \frac{3}{13}\epsilon. \tag{5}$$

Thus, a parameter is selected only when its empirical forget-domain sensitivity exceeds approximately 1.86 times its retain-domain sensitivity. Let $M = \text{diag}(m_i)$ denote the binary mask and $\widetilde{F}_{\mathcal{D}} = \text{diag}(F_{\mathcal{D}}(i))$ denote the diagonal empirical Fisher matrix. Any update $\Delta\theta$ supported on the mask consequently satisfies

$$\Delta\theta^\top M \widetilde{F}_{\mathcal{D}_r} M \Delta\theta \leq \frac{7}{13} \Delta\theta^\top M \widetilde{F}_{\mathcal{D}_f} M \Delta\theta. \tag{6}$$

This establishes a local relative-sensitivity property: under the diagonal empirical Fisher metric, permitted updates have lower retain-domain than forget-domain sensitivity, establishing a rigorous baseline against which we compare our primary geometric objective.

3.3 Targeted Manifold Scattering: Scattering the Domain Geometry

Although the Fisher-guided NegGrad baseline provides a structural constraint to help preserve foundational knowledge, our empirical evaluations reveal a critical limitation in highly constrained, open-vocabulary regimes. Specifically, at low class fractions ($\gamma \leq 0.50$), standard gradient ascent does not specify a target configuration in representation space. Under the sparse Fisher mask and limited class coverage in the few shot setup, this objective achieves only weak forgetting. Our decomposition over $\mathcal{C}_{\text{seen}}$ and $\mathcal{C}_{\text{unseen}}$ shows that this limited forgetting transfers across class partitions. The principal weakness is insufficient forgetting strength rather than the failure to generalize to unseen classes. This motivates supplementing the cross-entropy objective with explicit geometric preferences to strengthen forgetting while ensuring that the retention is not damaged.

Conversely, baseline methods like ADU rely on Maximum Mean Discrepancy (MMD) to explicitly separate domain distributions via prompt tuning. However, at low γ, MMD computes a separation shift based entirely on the empirical means of the small subset of seen classes. This calculated global shift, $\Delta \propto \mu_{\text{f,seen}} - \mu_{\text{r,seen}}$, becomes intrinsically entangled with the specific semantic geometry of $\mathcal{C}_{seen}$ rather than representing a pure domain translation. When this biased affine shift is applied globally across the latent space, an unseen retain embedding $z_r \in \mathcal{C}_{\text{unseen}}$ is forcefully displaced

($z_r \rightarrow z_r + \Delta$). This implicitly drags the embedding out of its ground-truth semantic cluster, mathematically ensuring the catastrophic forgetting that severely degrades the model's performance even on the retain domain $\mathcal{D}_r$.

These contrasting failure modes illustrate that domain unlearning requires *localized manifold diffusion*. Instead of indiscriminately pushing forget embeddings to infinity (NegGrad) or shifting the entire space globally (ADU), we must actively scatter the stylistic representations of the forget domain into dense, unrelated regions of the retain space. To achieve this, we introduce Targeted Manifold Scattering (TMS), a continuous geometric objective taking inspiration from Direct Preference Optimization (DPO) Rafailov et al. [2023]. While hard-margin objectives provide no further optimization signal once their margin is satisfied, our objective uses a log-sigmoid formulation to provide graduated penalties for the retention and confusion constraints. The optimization pressure decreases smoothly as each constraint is increasingly satisfied, rather than vanishing immediately at its margin.

Let z_r and z_f denote the L2-normalized visual embeddings for the retain and forget domains, respectively, with corresponding labels y_r and y_f. As overviewed in Figure 1, rather than treating all forget embeddings as uniform negatives, our TMS objective induces structural tension using two decoupled rules:

1. The Retention Component ($\mathcal{L}_{\text{retain}}$): This term protects the retain clusters from being encroached upon by same-class forget data. An anchor retain embedding z_r is forced to explicitly prefer its own same-class retain positive (z_{r_pos}) over a forget embedding z_f of the same class.

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{(z_r, z_f)|y_r=y_f} \left[-\log \sigma \left(\beta \left(s(z_r, z_{\text{r_pos}}) - s(z_r, z_f) - m_{\text{retain}} \right) \right) \right] \tag{7}$$

where $s(\cdot, \cdot)$ denotes cosine similarity, β is the temperature scaling, and m_{retain} is the retention margin gap.

2. The Targeted Confusion Component ($\mathcal{L}_{\text{confuse}}$): This term encourages forget embeddings to align with adaptively selected cross-class retain neighbors. Utilizing hard mining, each forget embedding z_f dynamically seeks out the single cross-class retain embedding it is currently most similar to (i.e., the easiest to mistake). The model is explicitly penalized unless it pushes this cross-class similarity above a high confusion margin:

$$\mathcal{L}_{\text{confuse}} = \mathbb{E}_{z_f} \left[-\log \sigma \left(\beta \left(\max_{z_r | y_r \neq y_f} s(z_f, z_r) - m_{\text{confuse}} \right) \right) \right] \tag{8}$$

The selected neighbor can differ across samples and change during optimization, rather than imposing a fixed incorrect target class. The final continuous geometric preference loss is a weighted sum of these decoupled forces:

$$\mathcal{L}_{\text{TMS}} = w_{\text{retain}} \mathcal{L}_{\text{retain}} + w_{\text{confuse}} \mathcal{L}_{\text{confuse}} \tag{9}$$

This preference objective is integrated directly with the base NegGrad cross-entropy (CE) terms using a scaling weight λ_{pref}:

$$\mathcal{L}_{\text{total}} = (1 - \lambda_{\text{ce}}) \mathcal{L}_{\text{retain_CE}} - \lambda_{\text{ce}} \mathcal{L}_{\text{forget_CE}} + \lambda_{\text{pref}} \mathcal{L}_{\text{TMS}} \tag{10}$$

The synergy between $\mathcal{L}_{\text{TMS}}$ and the Fisher mask is central to our open-vocabulary domain unlearning approach. The hard-mining confusion term encourages forget embeddings to align with dynamically selected cross-class retain neighbors, while the retention term favors separation from their same-class retain neighborhoods and maintains within-retain class alignment. Together, these terms encourage adaptive cross-class manifold misalignment in the pre-projection visual representation space, with updates restricted to parameters selected by the Fisher mask. This disrupts correct-class alignment on $\mathcal{C}_{\text{seen}}$, with our evaluations showing that the resulting forgetting also transfers to $\mathcal{C}_{\text{unseen}}$, while the retention term discourages disruption of retain-domain representations. Our ablations support the complementary contributions of the retention and confusion terms, while additional representation-displacement analysis in the Appendix A.4.2 shows that TMS induces larger, more domain-aligned changes to forget representations and smaller changes to retain representations than Fisher-guided NegGrad, helping explain its stronger forgetting–retention trade-off.

Furthermore, this extreme geometric demand renders our framework exceptionally sample-efficient. Because the hard-mining objective extracts maximal structural unlearning signal from every individual image, TMS effectively eliminates the reliance on massive, statistically stable datasets typically required to compute distribution shifts. Consequently, this targeted approach explicitly unlocks rapid, high-performance few-shot domain unlearning, achieving profound erasure with only a fraction of the data exposure required by prior works.

4 Experiments

4.1 Experimental Setup

Datasets. We extensively evaluate our method across three standard benchmarks, each featuring four visually distinct domains: **Office-Home** Venkateswara et al. [2017] (65 classes; Art, Clipart, Product, Real-World), **Mini-DomainNet** Peng et al. [2019] (126 classes; Clipart, Painting, Real, Sketch), and **PACS** Li et al. [2017] (7 classes with severe stylistic shifts; Photo, Art Painting, Cartoon, Sketch).

Baselines. We benchmark our proposed Targeted Manifold Scattering (TMS) against two primary methods: 1) **ADU** Kawamura et al. [2025], a recent state-of-the-art global prompt-tuning approach for domain unlearning, and 2) **NegGrad+Fisher**, the rigorous parameter-isolation baseline formalized in Section 3.1, which applies empirical Fisher-guided gradient ascent directly to the model weights. Additional baselines such as SEMU Sendera et al. [2025], and SAM based unlearning Tang and Khanna [2025] are considered in the Appendix.

Evaluation Protocol. Following the setup defined in Section 3.1, we adopt a strict few-shot ($K = 8$) unlearning protocol using a class fraction $\gamma \in \{0.25, 0.50, 0.75, 1.0\}$. We evaluate on all classes ($\mathcal{C}_{seen} \cup \mathcal{C}_{unseen}$) to assess both domain erasure and semantic retention, and further benchmark zero-shot generalization robustness on ImageNet-1K Deng et al. [2009].

Implementation Details. We employ OpenCLIP (ViT-B/16 Dosovitskiy et al. [2020]) with a frozen text encoder, updating only the visual weights constrained by a Fisher sensitivity mask. Our TMS loss is applied to the pre-projection CLS-token space to ensure foundational structural updates. To prevent gradient imbalance during few-shot sampling, the retain and forget cross-entropy terms are dynamically scaled by their sample proportions ($\alpha_r = \frac{n_r}{n_r+n_f}$ and $\alpha_f = \frac{n_f}{n_r+n_f}$). The network is optimized using AdamW Loshchilov and Hutter [2017] with a learning rate of 1e-5 for 10 epochs. Additional details are provided in the Appendix.

Evaluation Metrics. We evaluate unlearning performance using **Retain Accuracy ($\mathcal{A}_r$)** (higher is better) and **Forget Accuracy ($\mathcal{A}_f$)** (lower is better). To quantify the overall trade-off between effective domain erasure and zero-shot semantic preservation, we report their **Harmonic Mean ($\mathcal{H}$)**. Specifically, $\mathcal{H}$ is computed between $\mathcal{A}_r$ and the erasure rate ($\mathcal{E} = 100 - \mathcal{A}_f$) as $\mathcal{H} = \frac{2 \cdot \mathcal{A}_r \cdot \mathcal{E}}{\mathcal{A}_r + \mathcal{E}}$.

Table 1: Unlearning performance across different class fractions (γ) with $|D_f| = 1$. **F**: Forget domain accuracy (%), **R**: Retain domain accuracy (%), **HM**: Harmonic Mean (%). Best values for each fraction and dataset are in **bold**.

Dataset	Method	$\gamma = 0.25$			$\gamma = 0.50$			$\gamma = 0.75$			$\gamma = 1.0$		
		F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑
OfficeHome	Baseline	59.55	**65.75**	50.09	51.58	**69.21**	56.98	37.03	**71.89**	67.13	31.27	77.89	73.02
	ADU	**21.68**	27.50	40.70	**28.21**	49.04	58.27	33.47	67.76	67.14	36.37	**80.36**	71.02
	Ours	53.33	63.22	**53.70**	39.49	68.36	**64.19**	**24.91**	68.40	**71.59**	**24.41**	75.44	**75.52**
DomainNet	Baseline	55.67	**77.72**	56.46	27.98	**77.08**	74.46	20.04	**77.24**	78.57	18.02	77.78	79.83
	ADU	**14.64**	24.44	38.01	**19.80**	49.25	61.02	23.49	67.33	71.63	27.86	**81.48**	76.53
	Ours	34.05	74.97	**70.17**	20.08	75.40	**77.59**	**14.60**	76.02	**80.44**	**14.44**	76.88	**80.99**
PACS	Baseline	78.64	90.02	34.52	57.75	**93.76**	58.25	59.40	**93.42**	56.60	41.01	95.60	72.96
	ADU	**64.97**	28.65	31.52	64.07	59.53	44.81	68.36	74.40	44.40	72.48	93.16	42.49
	Ours	76.085	**90.39**	**37.82**	**47.06**	93.49	**67.60**	**47.07**	93.24	**67.53**	**38.35**	**95.64**	**74.97**

4.2 Experimental Results

Overall Efficacy and the TMS Formulation: Table 1 demonstrates that our TMS formulation strictly achieves the highest Harmonic Mean (HM) across all datasets (OfficeHome, DomainNet, and PACS) and class fractions (γ). This superior performance stems directly from our dual-objective design: a **confusion component** that aggressively unlearns target domain features, and a **retain component** that explicitly anchors preserved knowledge.

Seen/Unseen Decomposition. Aggregate metrics alone do not isolate open-vocabulary behavior. Table 4 reports forget accuracy, retain accuracy, and HM separately on $\mathcal{C}_{\text{seen}}$ and $\mathcal{C}_{\text{unseen}}$ for DomainNet. The forget and retain support sets share the same seen classes; unseen classes contribute neither

Table 2: Ablation study on the DomainNet dataset comparing the individual components of our TMS loss formulation (Confuse-only, Retain-only, and Both). **F**: Forget domain accuracy (%), **R**: Retain domain accuracy (%), **HM**: Harmonic Mean (%). Best aggregate values are in **bold**.

Domain	γ	Confuse Only			Retain Only			Both Losses (Ours)		
		F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑
Clipart	0.50	19.68	77.25	78.75	20.63	**78.04**	78.70	**17.93**	76.82	**79.35**
	1.00	**16.03**	**78.10**	**80.93**	18.73	77.78	79.49	17.30	77.08	79.79
Painting	0.50	21.27	78.15	78.44	26.03	**80.37**	77.04	**17.46**	76.82	**79.57**
	1.00	17.78	78.94	80.55	23.33	**80.37**	78.47	**17.30**	78.88	**80.74**
Real	0.50	**31.27**	71.48	**70.08**	38.57	**72.86**	66.66	32.38	72.70	70.06
	1.00	23.81	**76.61**	76.40	18.73	74.23	77.59	**13.65**	73.75	**79.55**
Sketch	0.50	14.44	75.08	79.98	13.49	**76.98**	**81.47**	**12.53**	75.23	80.88
	1.00	11.59	76.46	82.00	10.79	**77.99**	83.22	**9.52**	77.77	**83.64**
DomainNet (Avg)	0.50	21.67	75.49	76.81	24.68	**77.06**	75.97	**20.06**	75.39	**77.60**
	1.00	17.30	77.53	79.97	17.90	**77.59**	79.69	**14.44**	76.84	**80.97**

Table 3: Unlearning performance on DomainNet for multiple domain forgetting ($|D_f| = 3$) across different class fractions (γ). **F**: Forget domain accuracy (%), **R**: Retain domain accuracy (%), **HM**: Harmonic Mean (%). Best values for each fraction are in **bold**.

Dataset	Method	$\gamma = 0.25$			$\gamma = 0.50$			$\gamma = 0.75$			$\gamma = 1.0$		
		F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑	F ↓	R ↑	HM ↑
DomainNet	Baseline	8.92	59.87	72.25	**4.54**	61.06	74.48	**3.26**	62.13	75.67	**3.33**	63.45	76.61
	ADU	9.97	24.88	38.99	13.70	47.61	61.37	14.85	64.92	73.67	16.79	**77.06**	**80.02**
	Ours	**7.11**	**60.19**	**73.05**	5.07	**62.45**	**75.34**	4.81	**66.10**	**78.02**	4.59	67.21	78.86

a forgetting nor a retention signal during unlearning. Evaluation images are disjoint from the support set, and predictions use the full class vocabulary. any configuration changes.

Our method achieves the highest seen- and unseen-class HM at every evaluated class fraction. At $\gamma = 0.25$, ADU achieves an unseen forget accuracy of 5.69%, but unseen retain accuracy simultaneously falls to 2.91%, yielding an unseen HM of only 5.64%. Our method achieves 34.20% unseen forget accuracy while maintaining 72.04% unseen retain accuracy, yielding an unseen HM of 68.78%. Thus, low forget accuracy on unseen classes alone is insufficient evidence of selective domain unlearning: it must be considered alongside retention on those same classes. Fisher-guided NegGrad exhibits the complementary limitation, preserving retain accuracy but achieving weaker forgetting; at $\gamma = 0.25$, its unseen forget accuracy is 56.06%. These results support transfer of forgetting beyond the unlearning classes while limiting damage to unseen-class recognition in the retain domains. Additional seen/unseen results on Office-Home are provided in Appendix A.4.1, showing a similar disparity between seen- and unseen-class retention under ADU.

Overcoming the Stability-Plasticity Dilemma: The results reveal a severe trade-off in existing methods. While ADU achieves strong forgetting at low fractions ($\gamma \leq 0.50$), it suffers catastrophic degradation on the retain set (e.g., DomainNet retain accuracy collapses to 24.45% at $\gamma = 0.25$, whereas ours maintains 74.97%). Conversely, the Baseline preserves retain knowledge but fails to sufficiently unlearn the target concepts. By leveraging our independent retain and confusion components, our method successfully decouples these objectives, resolving the dilemma to deliver optimal HM.

Robustness to Unlearning Scope (γ): As the unlearning scope expands to the full domain ($\gamma = 1.0$), our confusion component efficiently drives forget accuracies to the lowest (best) levels across all benchmarks (e.g., 14.44% on DomainNet, 24.61% on OfficeHome). Unlike ADU, which struggles with mode collapse at varying fractions, our formulation scales seamlessly, ensuring that the retain component stabilizes the network while the confusion component maximizes unlearning capacity.

Multi-Domain Unlearning: Table 3 shows that even in the multi-domain forgetting scenario, our TMS formulation demonstrates superior stability and balance compared to both the Baseline and ADU. While ADU suffers from a catastrophic collapse in retention at lower class fractions (e.g., 24.88% at $\gamma = 0.25$), our method maintains robust performance and achieves the highest Harmonic Mean for the majority of settings. These results confirm that the synergy between our retain and

Table 4: Seen/unseen decomposition on DomainNet with $K = 8$ support images per seen class per domain. F: forget accuracy; R: retain accuracy; HM: harmonic mean. All values are percentages. Best HM for each class partition and class fraction is in bold.

		F ↓		R ↑		HM ↑	
γ	Method	Seen	Unseen	Seen	Unseen	Seen	Unseen
0.25	Baseline	54.53	56.06	85.42	75.11	59.35	55.44
	ADU	40.94	5.69	87.71	2.91	70.59	5.64
	Ours	33.59	34.20	83.59	72.04	**74.02**	**68.78**
0.50	Baseline	31.43	24.52	81.19	72.96	74.35	74.20
	ADU	31.83	7.78	84.74	13.76	75.56	23.94
	Ours	22.54	17.62	79.58	71.22	**78.50**	**76.39**
0.75	Baseline	21.54	15.62	80.53	67.55	79.48	75.03
	ADU	28.09	10.00	82.73	22.08	76.94	35.46
	Ours	15.96	10.62	78.95	67.40	**81.42**	**76.84**

confusion components scales effectively to complex unlearning tasks, providing precise domain erasure while safeguarding foundational knowledge significantly better than existing approaches.

Table 5: Zero-shot generalization robustness on ImageNet-val and CIFAR-100 after unlearning specific DomainNet domains. Higher values indicate better preservation of foundational, open-vocabulary knowledge without catastrophic semantic forgetting. The original base model serves as the upper bound reference.

Forget Domain	**Method**	**ImageNet-val (%) ↑**	**CIFAR-100 (%) ↑**
Base Model (No Unlearning)		*63.38*	*61.99*
Clipart	Baseline	54.11	57.40
	ADU	36.73	49.37
	Ours	**57.33**	**57.50**
Painting	Baseline	**51.76**	50.45
	ADU	28.45	32.54
	Ours	49.30	**52.26**
Sketch	Baseline	46.87	52.49
	ADU	39.42	**57.01**
	Ours	**48.82**	56.21

Ablation Study: Loss Components. Table 2 isolates the effects of the *confuse* and *retain* components within our TMS loss on the DomainNet benchmark. When applied individually, a clear stability-plasticity trade-off emerges. The *confuse-only* loss generally promotes target forgetting but can destabilize retain accuracy, whereas the *retain-only* loss maximizes knowledge preservation (achieving the highest average Retain accuracy of 77.59% at $\gamma = 1.0$) but struggles to induce sufficient forgetting.

By unifying both objectives, our complete formulation explicitly resolves this trade-off. The synergistic combination of targeted domain confusion with explicit structural retention yields the lowest overall forget accuracies ($F = 20.06\%$ and 14.44%) while maintaining highly competitive retention. Consequently, our joint formulation achieves the highest average Harmonic Mean (HM $= 77.60\%$ and 80.97%) across both class fractions, confirming that both components are strictly necessary for optimal domain unlearning.

Few-Shot Sample Efficiency: Figure 3 evaluates performance under extreme few-shot constraints ($n \in \{1, 4, 8\}$ shots per class). Our method demonstrates significantly faster convergence, achieving highly effective unlearning with just 4 shots and strictly outperforming the peak 8-shot performance of both the baseline and ADU. For closed-domain unlearning ($\gamma = 1.0$), our 4-shot Harmonic Mean (80.80%) surpasses the 8-shot baseline (79.76%) and ADU (76.53%). This trend holds for open-vocabulary domain unlearning ($\gamma = 0.5$) as well, where our 4-shot accuracy (76.79%) decisively beats the 8-shot baseline (74.09%) and ADU (61.06%). Unlike competing methods that degrade sharply

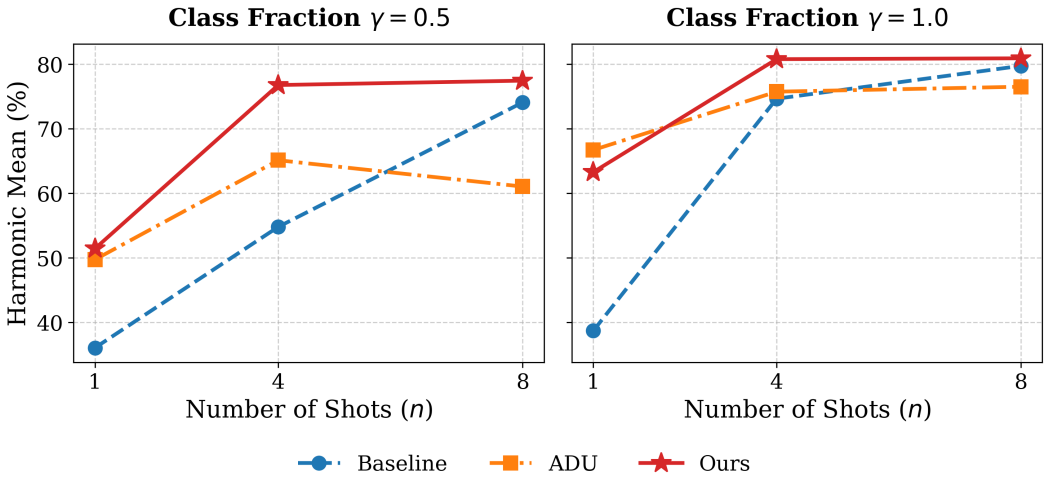

Figure 3: **Sensitivity to the Number of Training Samples.** Unlearning capacity (HM) compared across varying shots (n) per domain. While the Baseline and ADU show limited or unstable improvement in low-data regimes, our method demonstrates significantly faster convergence, achieving superior unlearning performance with just 4 shots.

in ultra-low data regimes ($n \leq 4$), our approach maintains robust feature decoupling, drastically reducing the data curation bottleneck required for practical unlearning.

Zero-Shot Generalization Robustness. A critical risk when unlearning specific visual domains (e.g., Painting or Sketch) is *catastrophic semantic forgetting*. To verify that our unlearning process is strictly localized and not damagingly broad, we benchmark zero-shot robustness on expansive, out-of-distribution datasets (ImageNet-val and CIFAR-100) following the domain erasure. As shown in Table 5, our TMS formulation excels at preserving global semantic knowledge. While prior methods like ADU suffer severe foundational degradation (e.g., dropping ImageNet accuracy by roughly 35% when unlearning the Painting domain), our method consistently anchors generalization performance near the original Base Model upper bound. Across all tested domains, our approach secures the highest overall retention on ImageNet and CIFAR-100 compared to the baselines. Crucially, this proves that direct parameter-tuning for unlearning—when formulated correctly—does not inherently destroy foundational generalization. Consequently, it demonstrates that relying on prompt-tuning-based unlearning is not strictly mandatory to safely avoid catastrophic forgetting in large vision-language models. Additional results can be found in the Appendix.

5 Conclusion and Limitations

We formalized the **Open-Vocabulary Domain Unlearning (OVDU)** protocol to address the hidden risk of catastrophic semantic forgetting in VLMs. By employing Fisher-guided parameter editing with our **Targeted Manifold Scattering (TMS)** objective, we demonstrate that domain stylistic geometry can be scattered without compromising foundational zero-shot intelligence. Our framework achieves superior stability-plasticity balance across 3 datasets, outperforming 8-shot baselines with only 4 samples. This confirms that constrained parameter-tuning is a robust and sample-efficient path for precise domain erasure. Despite these gains, several limitations remain. The Fisher mask provides a local selectivity prior: its diagonal approximation omits cross-parameter dependencies and does not guarantee global semantic preservation. Structured or low-rank Fisher approximations may better capture these dependencies. Additionally, while we prove scalability to $|D_f| = 3$, future work should investigate the upper bounds of multi-domain erasure and extend the TMS objective to the text encoder to prevent cross-modal information leakage.

Acknowledgments

This work was supported by Australian Research Council Project FT230100426 and partially supported by the use of the Nectar Research Cloud, a collaborative Australian research platform supported by the NCRIS-funded Australian Research Data Commons (ARDC).

References

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fartash Faghri, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, Alexander Toshev, Oncel Tuzel, and Hadi Pouransari. Mobileclip2: Improving multi-modal reinforced training. *arXiv preprint arXiv:2508.20691*, 2025.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Qian Feng, JiaHang Tu, Mintong Kang, Hanbin Zhao, Chao Zhang, and Hui Qian. Fg-oriu: Towards better forgetting via feature-gradient orthogonality for incremental unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1957–1967, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312, 2020.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11516–11524, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. *Advances in Neural Information Processing Systems*, 37:26377–26414, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. Approximate domain unlearning for vision-language models. *arXiv preprint arXiv:2510.08132*, 2025.

Hyo Seo Kim, Dongyoon Han, and Junsuk Choe. Negmerge: Sign-consensual weight merging for machine unlearning. In *International Conference on Machine Learning*, 2025.

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

Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research*, 2024.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Yusuke Kuwana, Yuta Goto, Takashi Shibata, and Go Irie. Black-box forgetting. *Advances in Neural Information Processing Systems*, 37:58792–58815, 2024.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Yufang Liu, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. Unlearning with fisher masking. *arXiv preprint arXiv:2310.05331*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Rymarczyk. Semu: Singular value decomposition for efficient machine unlearning. *arXiv preprint arXiv:2502.07587*, 2025.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Haoran Tang and Rajiv Khanna. Sharpness-aware machine unlearning. *arXiv preprint arXiv:2506.13715*, 2025.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in neural information processing systems*, 34:9706–9719, 2021.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

A Appendix

A.1 Extended Related Work

A.1.1 Machine Unlearning: From Classes to Domains

Machine unlearning aims to remove the influence of specific training data from a model without the prohibitive computational cost of retraining from scratch Bourtoule et al. [2021], Graves et al. [2021]. In the context of vision models, the literature has predominantly focused on *class unlearning* techniques designed to make the model forget specific object categories while preserving accuracy on retained classes Fan et al. [2023], Golatkar et al. [2020], Ilharco et al. [2022]. Within this realm of class unlearning, parameter isolation and direct weight editing have emerged as crucial strategies to prevent the catastrophic forgetting of retained knowledge. Recent advancements manipulate the parameter space directly to induce forgetting, employing techniques such as gradient weight negation Graves et al. [2021], subtracting task-specific vectors via Task Arithmetic Ilharco et al. [2022], or applying consensual weight merging as seen in NegMerge Kim et al. [2025]. Other approaches target specific representational subspaces, leveraging Singular Value Decomposition (SEMU) Sendera et al. [2025] or gradient-based weight saliency Fan et al. [2023] to mask and alter only a fraction of critical parameters. Similarly, Fisher-based approaches utilize empirical Fisher Information to identify and isolate the specific subset of weights that are disproportionately sensitive to the forget data Liu et al. [2023].

While these advanced weight-editing and isolation techniques have proven highly effective at preserving semantic boundaries during class erasure, their potential has remained entirely unexplored in the context of domain unlearning. Here, the objective shifts fundamentally: rather than deleting a discrete concept, the model must decouple a pervasive stylistic rendering from an open-vocabulary semantic space. Furthermore, as illustrated in Figure 2, the latent feature space of pre-trained VLMs exhibits significant entanglement among domain distributions, in stark contrast to typically well-separated class distributions. This severe overlap makes the successful application of conventional class unlearning techniques Kodge et al. [2024], Sendera et al. [2025], Feng et al. [2025] inherently challenging for domain unlearning.

Class-level erasure is often too blunt for practical applications, prompting the recent introduction of Approximate Domain Unlearning (ADU) Kawamura et al. [2025]. ADU aims to suppress a model's recognition capability for specific domains (e.g., illustrations) while retaining it for others (e.g., real photos). To address this, current ADU methods rely on prompt tuning combined with Maximum Mean Discrepancy (MMD) to explicitly disentangle domain distributions. However, as we demonstrate, these methods are evaluated in a closed-vocabulary setting where the unlearning and testing classes overlap perfectly. When applied to an open-vocabulary setting, prompt-based MMD induces a $\mathcal{C}_{train}$-biased global affine shift that severely degrades the zero-shot generalization manifold for unseen classes.

A.1.2 Domain Generalization and Unlearning in VLMs

While foundation models Faghri et al. [2025], Radford et al. [2021] exhibit profound domain generalization, their highly entangled latent spaces complicate unlearning. This often leads prior ADU works to rely on global separation techniques like MMD Gretton et al. [2012]. However, we observe that the *internal parameter space* of the visual encoder remains highly structured. Instead of forcing global feature disentanglement, our work leverages a Fisher Information Mask Liu et al. [2023] to isolate and degrade specific stylistic weights. This surgical parameter updating prevents the catastrophic semantic interference seen in global tuning approaches.

A.1.3 Geometric Regularization for Unlearning

Rigid metric learning objectives (e.g., Triplet Loss Hoffer and Ailon [2015]) are ill-suited for unlearning, as their hard margins typically cause simple *manifold translation* (relocating clusters) rather than actual feature disentanglement. To overcome this, we draw inspiration from Direct Preference Optimization (DPO) Rafailov et al. [2023] and its recent unlearning extensions like NPO Zhang et al. [2024]. We transition preference-based unlearning from token-level probabilities to continuous geometric representation editing via our *Targeted Manifold Scattering* (TMS) objective. By using hard mining to force forget embeddings to actively mimic unrelated cross-class retain

embeddings, TMS induces a localized *manifold diffusion* that scatters the stylistic geometry of the forget domain without compromising open-vocabulary generalization.

A.2 Extended Implementation Details

Optimization and Training Regime. The model is optimized using AdamW for 10 epochs with a learning rate of 1×10^{-5}, a weight decay of 1×10^{-2}, and a batch size of 64. A fresh optimizer is instantiated for every sweep configuration, and each run initializes from the same pre-trained CLIP weights. Experiments were predominantly conducted using NVIDIA A100 GPUs with 40 GB VRAM.

TMS Hyperparameters. For the geometric preference objective, we fix the temperature scaling at $\beta = 0.1$, the retention margin at $m_{\text{retain}} = 0$, and use equal weights for the retention and confusion components, $w_{\text{retain}} = w_{\text{confuse}} = 1$. The preference scaling weight λ_{pref} and confusion margin m_{confuse} are validated through grid search over $\lambda_{\text{pref}} \in \{0.5, 1.0, 2.0, 5.0\}$ and $m_{\text{confuse}} \in \{0.3, 0.5, 0.75\}$. For the TMS objective, the hard negatives are mined within each minibatch by selecting, for each forget sample, the retain sample with the highest cosine similarity among those with a different class label. Mining uses current model embeddings without a memory bank or queue; the confusion objective pulls each forget sample toward its selected neighbour.

Fisher Mask Computation. The diagonal empirical Fisher is computed once before unlearning using 10 minibatches from the respective domain data loaders. The domain-selectivity threshold is selected from $\tau \in \{0.15, 0.3, 0.5, 0.7\}$ using a small validation subset drawn exclusively from the training data; test samples are not used for threshold selection. For simultaneous multi-domain unlearning, samples from all target domains are combined and the model undergoes a single unlearning stage. We construct a shared mask using a rank-matched procedure: mask sizes obtained from the validated single-domain thresholds are averaged, the sensitivity ratio is recomputed on the combined forget set against the remaining retain domain, and the top-k parameters are selected.

Class Partitions and Support Counts. The class fraction γ determines the number of seen classes, while K specifies the number of support images per seen class per domain. Office-Home, Mini-DomainNet, and PACS contain 65, 126, and 7 classes, respectively, with four domains in each benchmark. Table 6 lists the resulting partitions and support sizes for $K = 8$. For example, Office-Home at $\gamma = 0.25$ uses 16 of its 65 classes for unlearning, with the remaining 49 classes held out. The support set contains 128 images from the forget domain and 384 images from the three retain domains, giving 512 images in total. No images or supervision from the 49 held-out classes enter unlearning. Similarly, DomainNet at $\gamma = 0.50$ uses 63 seen and 63 unseen classes, with 504 forget-domain and 1512 retain-domain support images.

Table 6: Class partitions and support-image counts for single-domain unlearning with $K = 8$ images per seen class per domain. Each dataset has four domains: one forget domain and three retain domains. Retain counts are totals across the three retain domains.

Dataset	γ	Seen	Unseen	Forget	Retain	Total
Office-Home	0.25	16	49	128	384	512
	0.50	32	33	256	768	1024
	0.75	49	16	392	1176	1568
	1.00	65	0	520	1560	2080
Mini-DomainNet	0.25	32	94	256	768	1024
	0.50	63	63	504	1512	2016
	0.75	94	32	752	2256	3008
	1.00	126	0	1008	3024	4032
PACS	0.25	2	5	16	48	64
	0.50	4	3	32	96	128
	0.75	5	2	40	120	160
	1.00	7	0	56	168	224

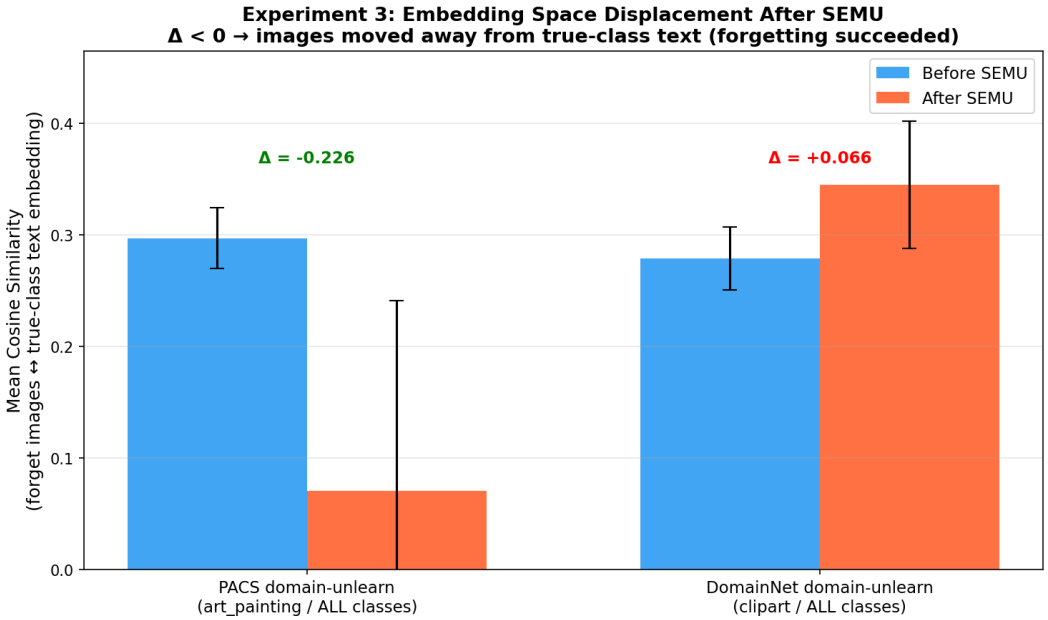

Figure 4: SEMU unlearning performance on PACS art_painting domain and the Domainnet clipart domain.

PACS Data Split. For PACS, we construct an 80% training and 20% test split using stratified sampling with a fixed random seed. The few-shot support images are sampled from the training partition, while evaluation uses the held-out test partition. The same split was used across methods.

Experimental runs were conducted multiple times while varying the random seed, class partition, and few-shot support samples. Across the evaluated domains, datasets, and class fractions, the maximum observed deviation in HM across repeated runs is at most 1.5%. This indicates that the measured forgetting–retention trade-off remains relatively stable under variations in class exposure, support sampling, and random initialization of the experimental procedure.

Metric Aggregation. For each dataset and class fraction, forget and retain accuracies are averaged across forget-domain configurations. The aggregate HM is then computed from these averaged accuracies. The same procedure is applied separately to the seen and unseen class partitions.

A.3 Additional Baselines

For additional baselines, we considered another recent unlearning method tested mostly on class unlearning. We adopt SEMU Sendera et al. [2025] as a primary baseline, leveraging its SVD-based projection method for efficient, data-free machine unlearning. But since it relies mostly on having distinct subspaces, the domain entanglement in pre-trained VLMs make it extremely challenging for such methods to work off the shelf. Results on domainnet dataset, especially, were worse than baselines we consider in the main paper. On PACS, though inferior to the methods in the main paper, it did perform unlearning. Figure 2 clearly shows that the domains in the domainnet dataset are very entangled compared to the PACS dataset making SEMU approach fail.

A.4 Additional Experiments

To test the scalability of our method, we tested it on a stronger CLIP-based VLM with a ViT-L-14 backbone and compared it with the baseline and ADU. Tables 7, 8, and 9 results clearly indicate that the proposed TMS objective improves forgetting by a large margin by conceding a very small margin in the retention, thereby achieving the best overall trade-off scores across the board.

A.4.1 Additional Seen/Unseen Decomposition

Table 10 extends the class-partitioned evaluation to Office-Home. Following the main-paper protocol, the forget and retain support sets share $\mathcal{C}_{\text{seen}}$, while $\mathcal{C}_{\text{unseen}}$ is absent from both. Evaluation images

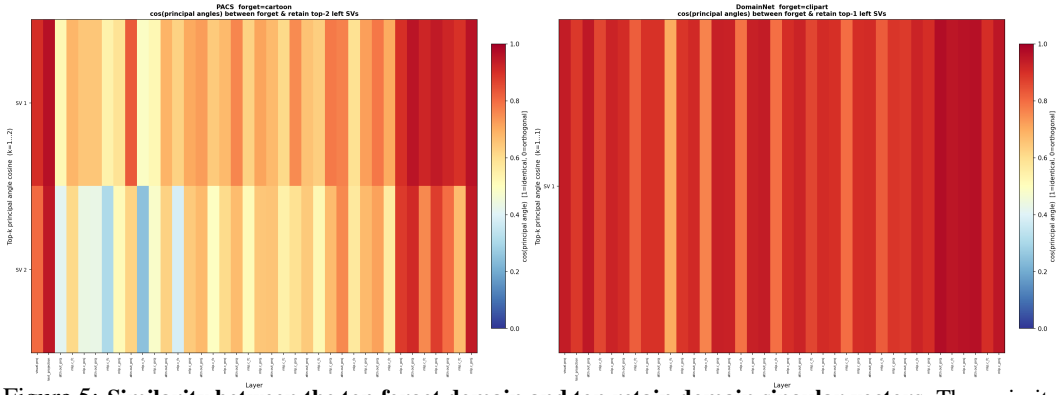

Figure 5: **Similarity between the top forget domain and top retain domain singular vectors.** The majority red amongst the domainnet domains implies that the orthogonality between domains do not exist making SEMU-based approaches very hard to get it working off the shelf.

Table 7: Comparison of Baseline and TMS Methods on a stronger pretrained ViT-L-14 backbone ($N = 8, \gamma = 0.5$).

Domain	**Forget Accuracy (↓)**		**Retain Accuracy (↑)**		**Harmonic Mean (↑)**	
	Baseline	TMS	Baseline	TMS	Baseline	TMS
Painting	0.3016	**0.2222**	**0.8603**	0.8254	0.7710	**0.8009**
Real	0.8651	**0.2937**	**0.8386**	0.8069	0.2324	**0.7533**
Clipart	0.2714	**0.1508**	**0.8354**	0.8164	0.7784	**0.8325**
Sketch	0.1635	**0.0841**	**0.8228**	0.8074	0.8296	**0.8582**

Table 8: Comparison of Baseline and TMS on a stronger pretrained ViT-L-14 backbone($N = 8$, $\gamma = 1.0$).

Domain	**Forget Accuracy (↓)**		**Retain Accuracy (↑)**		**Harmonic Mean (↑)**	
	Baseline	TMS	Baseline	TMS	Baseline	TMS
Painting	0.2222	**0.1714**	**0.8566**	0.8381	0.8153	**0.8333**
Real	0.1460	**0.1206**	0.8063	**0.8095**	0.8295	**0.8430**
Clipart	0.2079	**0.1349**	**0.8360**	0.8190	0.8134	**0.8414**
Sketch	0.1222	**0.0746**	**0.8349**	0.8212	0.8558	**0.8702**

Table 9: Comparison of Methods Across Domains ($N = 8$, $\gamma = 1.0$). Metrics are F: Forget Accuracy, R: Retain Accuracy, and HM: Harmonic Mean.

Method	**Painting**			**Real**			**Clipart**			**Sketch**		
	F (↓)	**R (↑)**	**HM (↑)**	**F (↓)**	**R (↑)**	**HM (↑)**	**F (↓)**	**R (↑)**	**HM (↑)**	**F (↓)**	**R (↑)**	**HM (↑)**
Baseline	0.2222	0.8566	0.8153	0.1460	0.8063	0.8295	0.2079	0.8360	0.8134	0.1222	0.8349	0.8558
ADU	0.2524	**0.8815**	0.8091	0.2683	**0.8503**	0.7866	0.2365	**0.8624**	0.8099	0.2365	**0.8847**	0.8196
TMS (Ours)	**0.1714**	0.8381	**0.8333**	**0.1206**	0.8095	**0.8430**	**0.1349**	0.8190	**0.8414**	**0.0746**	0.8212	**0.8702**

are disjoint from the support set, and all predictions use the full vocabulary of 65 classes. All ADU results were reproduced using the authors' official codebase[1] without any configuration changes.

Retention on unseen classes. ADU exhibits a substantial disparity between seen- and unseen-class retention. At $\gamma = 0.25$, its seen retain accuracy is 88.16%, whereas unseen retain accuracy falls to 7.80%. Consequently, despite an unseen forget accuracy of 12.26%, its unseen HM is only 14.32%. Our method maintains 57.39% unseen retain accuracy and achieves an unseen HM of 52.06%. A similar pattern holds at $\gamma = 0.50$: ADU retains 83.90% accuracy on seen classes but only 12.90% on unseen classes, compared with 77.50% and 60.14%, respectively, for our method. These results reinforce the importance of evaluating retention on the same unseen classes used to assess open-vocabulary forgetting.

[1] `https://github.com/kodaikawamura/domain-unlearning`

Table 10: Seen/unseen decomposition on Office-Home with $K = 8$ support images per seen class per domain. F: forget accuracy; R: retain accuracy; HM: harmonic mean. All values are percentages. Best HM for each class partition and class fraction is in bold.

		F ↓		R ↑		HM ↑	
γ	Method	Seen	Unseen	Seen	Unseen	Seen	Unseen
0.25	Baseline	62.51	58.61	81.72	60.60	51.40	49.19
	ADU	51.67	12.26	88.16	7.80	**62.43**	14.32
	Ours	56.38	52.36	81.32	57.39	56.78	**52.06**
0.50	Baseline	50.55	52.50	78.27	61.06	60.61	53.43
	ADU	40.27	15.84	83.90	12.90	**69.78**	22.37
	Ours	38.14	40.67	77.50	60.14	68.80	**59.74**
0.75	Baseline	38.58	32.58	79.45	49.87	69.28	**57.33**
	ADU	37.49	20.81	80.00	28.75	70.18	42.18
	Ours	25.72	22.63	76.49	44.83	**75.37**	56.77

Forgetting–retention trade-off. Our method achieves higher unseen-class HM than ADU at every evaluated class fraction, although ADU achieves higher seen-class HM at $\gamma = 0.25$ and $\gamma = 0.50$. Compared with Fisher-guided NegGrad, our method achieves higher unseen-class HM at these two fractions. At $\gamma = 0.75$, the Baseline has a marginally higher unseen-class HM (57.33% versus 56.77%), while our method provides substantially stronger seen-class forgetting (25.72% versus 38.58% forget accuracy). Considering all classes together, our method achieves the highest overall HM at every evaluated fraction, as reported in Table 1.

Choice of benchmarks. We use DomainNet and Office-Home as the primary benchmarks for class-partitioned analysis because they contain 126 and 65 classes, respectively. PACS contains only seven classes, yielding seen/unseen partitions of 2/5, 4/3, and 5/2 classes at the evaluated fractions. Its partitioned metrics are therefore more sensitive to the behavior of individual classes. We retain PACS as a complementary low-data benchmark in the aggregate evaluation.

A.4.2 Diagnostic Analysis of Representation Updates

To investigate why TMS improves upon Fisher-guided NegGrad, we compare their representation changes on DomainNet with $\gamma = 0.25$ and Painting as the forget domain. Both methods use the Fisher-masked cross-entropy objective; our method additionally includes the retention and confusion preferences. For each evaluation image, we measure $d(x) = z_{\text{method}}(x) - z_{\text{base}}(x)$, using image embeddings after projection and before L2 normalization.

Diagnostic definitions. We estimate a fixed domain direction from base-model embeddings of test images belonging to $\mathcal{C}_{\text{seen}}$. Let μ_d denote their mean embedding in domain d, and let μ_R be the equally weighted average of the three retain-domain means. The unit direction is

$$\hat{\delta} = \frac{\mu_{\text{painting}} - \mu_R}{\|\mu_{\text{painting}} - \mu_R\|_2}. \tag{11}$$

This direction is reused for both methods and all evaluation groups, solely for post-hoc analysis. For each group, we report mean displacement magnitude and the fraction of displacement energy along this axis:

$$\text{Expl.} = \frac{\sum_d w_d \,\text{mean}_{i\in d}\big[(d_i^\top \hat{\delta})^2\big]}{\sum_d w_d \,\text{mean}_{i\in d}\big[\|d_i\|_2^2\big]}. \tag{12}$$

where domains within the group receive equal weights. This is a ratio of averaged energies and measures axial concentration of displacement irrespective of its sign.

We also compute the correct-prototype margin

$$m(x) = \cos(z(x), t_y) - \max_{c\neq y} \cos(z(x), t_c), \tag{13}$$

using fixed text prototypes from the frozen base text encoder with the basic template with class prompts. Competitors span the full class vocabulary, including when evaluating a seen-only or

unseen-only group. The Lost metric is the fraction of all images in a group whose margin changes strictly from positive under the base model to negative after unlearning. All diagnostics are reported separately for forget and retain domains and for $\mathcal{C}_{\text{seen}}$ and $\mathcal{C}_{\text{unseen}}$.

Table 11: Representation-update diagnostics on DomainNet with $\gamma = 0.25$ and Painting as the forget domain. $\|d\|$: mean representation-displacement magnitude; Expl.: fraction of displacement energy along $\hat{\delta}$; Lost: fraction of all group images whose correct-prototype margin changes from strictly positive to strictly negative. Expl. and Lost are reported as fractions, not percentages.

		Forget		Retain	
Metric	Method	Seen	Unseen	Seen	Unseen
$\|d\|$	Baseline	6.47	6.57	6.16	6.16
	Ours	8.03	7.90	5.86	5.84
Expl.	Baseline	0.088	0.091	0.044	0.046
	Ours	0.143	0.144	0.053	0.053
Lost	Baseline	0.206	0.181	0.010	0.055
	Ours	0.425	0.377	0.029	0.067

More selective representation changes. As shown in Table 11, the Baseline produces similar displacement magnitudes on forget and retain images. For seen classes, their ratio is approximately $6.47/6.16 = 1.05$, compared with $8.03/5.86 = 1.37$ for our method. TMS therefore produces larger changes to forget representations while reducing the magnitude of changes to retain representations. The same pattern holds for unseen classes: forget displacement increases from 6.57 to 7.90, whereas retain displacement decreases from 6.16 to 5.84.

Domain-axis concentration and recognition changes. The fraction of forget-side displacement energy along the domain axis increases from 0.088 to 0.143 for seen classes and from 0.091 to 0.144 for unseen classes. Retain-side fractions remain comparatively low, although they increase slightly under TMS. These values indicate greater concentration along the domain axis, without specifying the direction of movement. The fraction of forget images losing their correct-prototype margin increases from 20.6% to 42.5% on seen classes and from 18.1% to 37.7% on unseen classes. The corresponding retain-side fractions remain below 7%, although they are higher than those of the Baseline. Together, these measurements show larger forget-side changes, smaller retain-side displacement, and stronger suppression of correct recognition on both class partitions, helping explain the improved forgetting–retention trade-off.

A.5 Computational Efficiency

Despite leveraging the empirical Fisher Information Matrix—a technique often associated with prohibitive computational overhead—our framework is designed for rapid deployment. Rather than continuously re-evaluating parameter sensitivity, the Fisher mask is computed strictly once prior to the unlearning phase using a minimal subset of data (10 minibatches). Furthermore, the structural tension induced by our Targeted Manifold Scattering (TMS) objective extracts maximal unlearning signal from every individual image via dynamic hard-mining. This geometric efficiency drives exceptionally fast convergence, allowing the entire domain erasure process to complete in just 10 epochs on a highly constrained few-shot dataset ($K = 8$). Combined with fully frozen, pre-computed text features and sparse updates restricted to the visual encoder, our approach achieves open-vocabulary domain unlearning at a fraction of the computational cost required by standard retraining or continuous prompt-tuning baselines.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: There are no theoretical results.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The codebase will be made open-source upon acceptance.

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are averaged across multiple runs to ensure stability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our framework enhances the safety of mission-critical systems, such as autonomous driving, by preventing hazardous out-of-distribution misclassifications. Furthermore, the extreme computational efficiency of our method significantly reduces the carbon footprint associated with model alignment, democratizing access to safe AI customization. However, like any targeted editing technique, it carries dual-use risks; malicious actors could potentially induce intentional blindspots in open-source models. Future work must focus on developing robust unlearning verification tools to mitigate these risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: Does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The code will be released upon acceptance.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing nor research with humans.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM used only for editing the paper for better communication.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.