FORGET AT SIGHT: TRAINING-FREE DOMAIN UNLEARNING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models such as CLIP are designed to recognize the same object across different visual domains. This makes domain unlearning difficult: an intervention that suppresses recognition in one domain can also damage the same classes in other domains. Existing approaches rely on iterative optimization and access to retain-domain images, limiting their use when model updates are costly or retain data are unavailable. We introduce GATE-DU (Gated Activation Steering for Targeted Erasure in Domain Unlearning), a training-free framework that separates *when to erase* from *how to erase*. For each input, a domain gate first estimates whether it belongs to the forget domain. Only then does GATE-DU contract its representation toward a class-agnostic forget-domain centroid, suppressing class recognition while leaving retain-domain inputs unchanged. GATE-DU obtains its gating evidence in three ways. When retain images are unavailable, GATE-DU (T) compares the forget-domain description against known retain-domain descriptions using the frozen text branch. GATE-DU (O) starts from this text gate and progressively grounds it in visual evidence from the unlabeled evaluation stream, first through a class-paired prototype and then through domain-LDA. When labeled retain images are available, GATE-DU (V) directly constructs a class-paired domain-LDA gate and calibrates it using a closed-form Gaussian posterior. All variants keep the model, prompts, and erasure payload frozen, requiring no gradients, adapters, or replay. Experiments on PACS, OfficeHome, and MiniDomainNet across class fractions $\gamma \in \{0.25, 0.5, 0.75, 1.0\}$ show that GATE-DU is competitive with or outperforms optimization-based methods, particularly in low-γ open-vocabulary settings. Our results highlight conditional intervention as a practical principle for domain unlearning: effective erasure requires deciding not only how to forget, but also when to forget.

1 INTRODUCTION

Machine unlearning (Bourtoule et al., 2021) aims to remove a specified capability, concept, or data influence from a trained model while preserving its remaining utility. Existing work has studied this problem extensively for discriminative vision models, most commonly through sample- and class-level removal, as well as for generative and diffusion models (Golatkar et al., 2020; Fan et al., 2024; Kodge et al., 2024; Wu & Harandi, 2025; Lu et al., 2024). In large language models, closely related objectives arise in safety, alignment, and knowledge removal, motivating both optimization-based and inference-time interventions (Zhang et al., 2024; Choi & Na, 2023; Sheng et al., 2026; Rimsky et al., 2024). These approaches serve two practically distinct deployment protocols. When model weights are to be released, a weight-level intervention may be necessary because the recipient controls the model and its inference pipeline. In a hosted or otherwise controlled deployment, however, the provider retains control over both weights and inference, and the operational objective is to suppress a specified behavior reliably at serving time. This second protocol is a *deployment-time functional unlearning* that avoids repeated optimization and can be applied to a frozen model. Hosted APIs and controlled enterprise inference pipelines make this setting practically important, particularly when unlearning requests are frequent, support data are limited, or retraining is costly.

We study this problem for *domain unlearning* in vision–language models (VLMs) (Kawamura et al., 2026). Unlike class unlearning, which asks a model to suppress a semantic category everywhere, domain unlearning asks it to stop recognizing all classes within a target visual domain while continuing

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.

to recognize the same classes in the remaining domains. For example, the model may be required to forget paintings while preserving its recognition of the same objects in photographs, sketches, and clipart. Open-vocabulary domain unlearning (OVDU) makes the problem harder: only a fraction of classes is available during unlearning, but forgetting and retention are evaluated over the complete class inventory, including unseen classes (Udupa et al., 2026). A successful method must therefore remove a domain-level behavior without memorizing the observed class–domain pairs.

This requirement exposes a central difficulty. VLMs are trained to preserve class semantics across visual domains, so the class-discriminative representations of the forget and retain domains are strongly entangled. Consequently, an unconditional edit that damages class evidence in the forget domain also damages the same class evidence in retain domains. Our analysis further shows that directly removing a putative forget-domain subspace offers little protection when domain and class directions overlap and different domains share substantial visual structure. Prior domain-unlearning methods address this problem through encoder fine-tuning, prompt adaptation, parameter isolation, or preference-based representation scattering (Kawamura et al., 2026; Udupa et al., 2026). Although effective when sufficient labeled support is available, these methods require gradient updates and typically rely on retain-domain images to constrain collateral damage. These requirements increase cost and can make the methods brittle in the few-shot, open-vocabulary regime for which both class coverage and retain supervision are limited.

We therefore introduce **GATE-DU** (**G**ated **A**ctivation steering for **T**argeted **E**rasure in **D**omain **U**nlearning), a conditional framework that separates deciding *when to forget* from executing the forgetting intervention. For each image, a domain gate determines the strength of a class-agnostic centroid-collapse payload. Forget-domain representations are contracted to suppress their class structure, including for unseen classes, while retain-domain inputs are left unchanged. In the most restrictive setting, where only labeled forget support and domain descriptions are available, **GATE-DU (T)** constructs this gate from the frozen VLM text branch. It contrasts class-conditioned domain prompts with neutral prompts, marginalizes the resulting evidence over the model's soft class predictions, and activates the payload only when the forget-domain description outperforms every known retain-domain alternative. This provides retain-image-free domain unlearning without gradients, adapters, or replay.

GATE-DU (O) extends this text-only gate using the unlabeled evaluation stream. After each prediction is recorded, the frozen text gate admits confident pseudo-retain samples, while CLIP provides their pseudo-classes. These samples update memory-efficient visual statistics without storing images or activations. A signed, class-balanced prototype first introduces visual evidence in the low-sample regime; as the statistics mature, it is refined into a covariance-aware domain-LDA gate. Reliability-controlled transitions determine how much the system trusts the text, prototype, and LDA evidence. The protocol remains strictly causal, with each update affecting only future samples.

When labeled retain images are available, as assumed by prior domain-unlearning methods, **GATE-DU (V)** initializes the gate directly from visual evidence. Class-paired statistics reduce the influence of class content, a regularized generalized eigenproblem selects the domain-discriminative layer and subspace, and a shared-covariance Gaussian posterior calibrates the forget-domain evidence. GATE-DU (V) therefore provides a fully closed-form alternative to optimization-based unlearning under the same supervision.

Together, the three variants form a unified training-free framework. GATE-DU (T) provides an immediate text-based gate without retain images, GATE-DU (O) progressively grounds this gate using visual evidence from the unlabeled stream, and GATE-DU (V) constructs a fully visual gate when labeled retain support is available. Across all three variants, the VLM and steering payload remain frozen; only the source and maturity of the gating evidence change.

Our contributions are four-fold:

- We introduce **GATE-DU (T)**, a retain-image-free and training-free domain gate for VLMs. It uses text descriptions to separate domain evidence from class information, compares the forget domain against the known retain domains, and applies a class-agnostic centroid-collapse payload only to images identified as belonging to the forget domain.
- We develop **GATE-DU (O)**, which extends the text gate through causal online visual grounding. The frozen text gate identifies confident pseudo-retain samples from the un-

labeled stream, which are summarized using running class-wise statistics without storing images or activations. These statistics first construct a visual prototype and, once sufficient stable evidence is available, a domain-LDA gate. Each image is evaluated before it updates the gate, and the procedure requires neither gradients nor evaluation labels.

- When labeled retain support is available, we introduce **GATE-DU (V)**, a closed-form, retain-initialized domain-LDA Gaussian gate. Class-paired forget–retain differences reduce the influence of class identity, the LDA objective selects the most domain-discriminative layer and subspace, and a Gaussian posterior provides a soft gate without iterative training.
- We provide a **comprehensive closed and open-vocabulary evaluation** of GATE-DU on Mini-DomainNet, OfficeHome, and PACS across different exposed class fractions and forget domains. The proposed training-free variants remain competitive with, and often outperform, existing optimization-based methods, particularly in low-class-fraction settings where limited support makes gradient-based unlearning difficult.

2 Related Works

Machine Unlearning. Machine unlearning removes designated samples, classes, or concepts while preserving retained utility. Vision methods largely address sample- or class-level deletion, whereas language-model unlearning targets memorized, harmful, or copyrighted content. NegGrad maximizes forget-data loss, while NPO regularizes this objective against a reference model to limit utility collapse (Choi & Na, 2023; Zhang et al., 2024). SalUn restricts updates using gradient saliency, NegMerge removes shared components of unlearning task vectors, and SEMU confines updates to an SVD-derived subspace with retain-based and retain-free variants (Fan et al., 2024; Kim et al., 2025; Sendera et al., 2025). These methods modify parameters and assume that unwanted knowledge can be localized by sample, class, or parameter subspace. Domain unlearning instead targets visual renderings shared across classes and entangled with retained representations, making global weight edits or subspace removal prone to damaging shared class information.

Approximate Domain Unlearning. Approximate Domain Unlearning (ADU) formalizes the removal of recognition within a visual domain while preserving the remaining domains, and tunes instance-conditioned visual prompts using classification and distribution-alignment losses (Kawamura et al., 2026). Open-Vocabulary Domain Unlearning (OVDU) extends evaluation to classes unseen during unlearning and combines Fisher-guided parameter isolation with targeted manifold scattering (Udupa et al., 2026). Both require gradient optimization and retain-domain images. In contrast, we study *functional unlearning*: GATE-DU keeps the VLM and prompts frozen and conditionally applies an inference-time erasure payload. Its retain-image-free variant uses only forget support and domain descriptions, while its retain-initialized variant analytically constructs a visual gate, avoiding backpropagation and learned adapters.

Online Learning. Online test-time adaptation learns sequentially from unlabeled target streams, commonly through entropy minimization or temporally stabilized self-training (Wang et al., 2020; 2022). Recent VLM methods instead adapt predictions using retrieved test-time evidence (Fan et al., 2025). GATE-DU (O) follows a prequential protocol but updates only class-wise sufficient statistics for conditional unlearning, leaving the VLM parameters frozen.

Activation and Conditional Steering. Activation steering controls model behavior by modifying intermediate representations at inference. ActAdd and contrastive activation steering derive control directions from activation differences and inject them at selected layers (Turner et al., 2023; Rimsky et al., 2024). AlphaSteer learns input-dependent transformations that protect benign activations through a null-space constraint while redirecting malicious ones toward refusal (Sheng et al., 2026). Such subspace-based protection assumes separable targeted and protected activation geometry; in domain unlearning, however, forget- and retain-domain class subspaces are highly overlapping, so global steering either leaves forget information intact or damages shared class semantics.

Conditional steering instead separates *when to intervene* from *how to intervene* (Weng et al., 2026). Gated Subspace Steering (GSS), for example, applies a corrective steer only when a

learned probe detects memorization-relevant activations (Zhang et al., 2026). GATE-DU adopts this paradigm without gradient updates: domain-text margins provide a retain-image-free gate, class-paired domain-LDA with a Gaussian posterior provides a visual gate, and a class-agnostic centroid-collapse payload performs erasure. Retention is therefore protected by avoiding intervention on retain inputs rather than requiring a globally non-interfering steering subspace.

3 Proposed Approach

3.1 Problem Setting and GATE-DU Overview

We consider a frozen vision-language model with vision encoder f_v and text encoder f_t. Let d_f denote the domain to be forgotten, $\mathcal{D}_R$ the set of retained domains, and $\mathcal{C}$ the class vocabulary. Following the few-shot open-vocabulary protocol of **?**, unlearning observes K images from d_f for only a fraction γ of the classes, while evaluation covers both observed and held-out classes across the forget and retain domains.

We propose **GATE-DU** (**G**ated **A**ctivation steering for **T**argeted **E**rasure in **D**omain **U**nlearning), a unified framework that separates domain unlearning into two components: a gate that estimates whether the current image belongs to d_f, and a payload that suppresses recognition according to the gate activation. This separation is useful because class-discriminative features are largely shared across domains; applying the payload indiscriminately would therefore also damage retain-domain recognition.

GATE-DU supports three operating regimes that differ only in the source of gating evidence. **GATE-DU (T)** constructs a retain-image-free gate from domain-text evidence. **GATE-DU (O)** begins from this text gate and progressively grounds it using visual statistics obtained from an unlabeled stream, first through a low-sample prototype and later through a covariance-aware domain-LDA discriminator. When labeled retain images are available from the outset, **GATE-DU (V)** directly constructs the visual domain-LDA Gaussian gate. The vision encoder, text encoder, prompt embeddings, and forgetting payload remain frozen in all three variants.

3.2 GATE-DU (T): Retain-Image-Free Text Gate

GATE-DU (T) considers the setting in which labeled forget support and the names or descriptions of the retain domains are available, but no retain-domain image is provided. Since a visual forget-versus-retain discriminator cannot be estimated in this setting, the gate obtains its domain evidence from the aligned text branch of the CLIP (Radford et al., 2021) VLM model.

Let $\mathcal{P}$ denote the set of prompt templates. For class c, template p, and domain d, let $t^d_{p,c}$ and $t^0_{p,c}$ denote the normalized embeddings of the domain-conditioned and neutral prompts, respectively. We define $\bar{t}^0_c = |\mathcal{P}|^{-1} \sum_{p\in\mathcal{P}} t^0_{p,c}$ and $\delta^d_{p,c} = t^d_{p,c} - t^0_{p,c}$.

Let $\alpha = \exp(s_{\mathrm{VLM}}) = 1/\tau_{\mathrm{VLM}}$ denote the frozen logit-scaling factor of the pretrained VLM, where τ_{VLM} is its learned contrastive temperature. This value is inherited from the pretrained model and is not tuned during unlearning. Given the normalized, unsteered image embedding $z(x)$, its soft class weights are

$$w_c(x) = \frac{\exp\big(\alpha \, \langle z(x), \bar{t}^0_c \rangle\big)}{\sum_{c' \in \mathcal{C}} \exp\big(\alpha \, \langle z(x), \bar{t}^0_{c'} \rangle\big)}. \tag{1}$$

For domain d and template p, the class-marginalized domain evidence is $m_{d,p}(x) = \sum_{c\in\mathcal{C}} w_c(x) \, \langle z(x), \delta^d_{p,c} \rangle$.

We aggregate this evidence across prompt templates as

$$\bar{m}_d(x) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} m_{d,p}(x),$$

$$\sigma_d(x) = \sqrt{\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(m_{d,p}(x) - \bar{m}_d(x)\right)^2}, \tag{2}$$

$$s_d(x) = \bar{m}_d(x) - \eta_{\text{prompt}} \sigma_d(x).$$

This aggregation mildly penalizes domain predictions that are inconsistent across prompt templates. The final text margin compares the forget-domain evidence against the strongest retain-domain alternative:

$$m_{\text{text}}(x) = s_{d_f}(x) - \max_{d \in \mathcal{D}_R} s_d(x). \tag{3}$$

A class-disjoint subset of the forget support is used to estimate a threshold τ_{text} at the desired forget-support coverage. To provide a signed score with a stable scale, we define $\ell_{\text{text}}(x) = \big(m_{\text{text}}(x) - \tau_{\text{text}}\big)/s_{\text{text}}$, where s_{text} is a robust dispersion estimate computed from the calibration margins. A positive value denotes forget-side evidence. The calibration split is used only to construct this initial text boundary, while all provided labeled forget images may still contribute to the payload and subsequent visual statistics. This text gate is deployed directly in GATE-DU (T) and serves as the initialization and fixed admission rule for GATE-DU (O).

3.3 GATE-DU (O): Online Visual Grounding

Although GATE-DU (T) provides an immediate retain-image-free gate, textual domain descriptions may not completely capture the visual appearance of a dataset domain. GATE-DU (O) therefore allows the unlabeled evaluation stream to progressively supplement the text boundary with visual retain evidence. The protocol is prequential: each image is first evaluated using the current state, after which it may update the state used for future images.

Let $q(c \mid x_t)$ be the frozen CLIP class distribution and $\hat{c}_t = \arg\max_c q(c \mid x_t)$. A stream image is admitted as pseudo-retain only when it lies confidently on the retain side of the frozen GATE-DU (T) boundary and has sufficiently low normalized class entropy. Importantly, admission continues to use the fixed text gate even after the online visual components become active. This prevents a visual candidate from determining the pseudo-labels used to construct itself.

For every candidate layer and pseudo-class, admitted activations update a running count, mean, and within-cell scatter using Welford recurrences (Welford, 1962). Raw stream images and activations are not retained, keeping the online state memory-efficient. The labeled forget statistics remain fixed and provide the trusted side of each forget–retain comparison.

Class-paired visual prototype. Let ℓ_p denote the fixed prototype layer, and let $n_{f,c}^{\ell_p}$ and $n_{r,c,t}^{\ell_p}$ denote the numbers of forget and admitted pseudo-retain samples, respectively, for class c. The paired-class set at time t is

$$\mathcal{P}_t = \left\{ c \in \mathcal{C}_{\text{support}} \;\middle|\; n_{f,c}^{\ell_p} > 0 \;\wedge\; n_{r,c,t}^{\ell_p} > 0 \right\}. \tag{4}$$

For every $c \in \mathcal{P}_t$, we compute the forget–retain mean difference $\Delta_{c,t} = \mu_{f,c}^{\ell_p} - \mu_{r,c,t}^{\ell_p}$ and its normalized direction $\widehat{\Delta}_{c,t} = \Delta_{c,t}/\big(\|\Delta_{c,t}\|_2 + \epsilon_\Delta\big)$.

The class-balanced mean direction and the resulting unit prototype are

$$\bar{\Delta}_t = \frac{1}{|\mathcal{P}_t|} \sum_{c \in \mathcal{P}_t} \widehat{\Delta}_{c,t}, \qquad d_t = \frac{\bar{\Delta}_t}{\|\bar{\Delta}_t\|_2 + \epsilon_d}. \tag{5}$$

Finally, the prototype boundary is the class-balanced mean of the projected forget–retain midpoints:

$$b_t = \frac{1}{|\mathcal{P}_t|} \sum_{c \in \mathcal{P}_t} d_t^\top \left(\frac{\mu_{f,c}^{\ell_p} + \mu_{r,c,t}^{\ell_p}}{2} \right). \tag{6}$$

The resulting prototype evidence is $\ell_{\mathrm{proto},t}(x) = \left(d_t^\top h(x) - b_t\right)/s_{\mathrm{proto},t}$, where $s_{\mathrm{proto},t}$ is estimated from the projected within-cell dispersion. This stage uses only class-paired first-order visual statistics and can therefore contribute before a stable covariance estimate is available.

GATE-DU (O) combines the original text evidence and the online prototype evidence as

$$\ell_{\mathrm{pre},t}(x) = (1 - \rho_t)\ell_{\mathrm{text}}(x) + \rho_t \ell_{\mathrm{proto},t}(x), \tag{7}$$

where $\rho_t \in [0, 1]$ summarizes the reliability of the current prototype. It depends on the admitted retain mass, paired-class coverage, and the stability of the estimated direction and boundary. An invalid or unstable refresh cannot replace the last valid state.

3.4 Domain-LDA Visual Gate

The domain-LDA construction provides the mature visual component of GATE-DU (O) and the initialization of GATE-DU (V). In the online variant, it is constructed only after the running statistics contain sufficient pseudo-retain evidence. In the retain-initialized variant, the same construction is applied immediately using labeled forget and retain support.

For GATE-DU (O), let $\mathcal{P}_t^{\mathrm{LDA}}$ denote the set of classes satisfying the required forget and pseudo-retain sample counts at an LDA refresh. For each candidate layer ℓ, define

$$\begin{aligned} \Delta_{c,t}^{\ell} &= \mu_{f,c}^{\ell} - \mu_{r,c,t}^{\ell}, \qquad c \in \mathcal{P}_t^{\mathrm{LDA}}, \\ S_{B,t}^{\ell} &= \sum_{c \in \mathcal{P}_t^{\mathrm{LDA}}} \Delta_{c,t}^{\ell} \left(\Delta_{c,t}^{\ell}\right)^\top. \end{aligned} \tag{8}$$

Let S_W^ℓ denote the pooled within-domain, within-class scatter. Class pairing reduces the contribution of shared class semantics, while S_W^ℓ accounts for variation within each domain–class cell. We solve $S_B^\ell w_i = \lambda_i \left(S_W^\ell + \epsilon I\right) w_i$ and select the layer using $J(\ell) = \sum_{i=1}^{k} \lambda_i$, where k retains the dominant generalized-eigenvalue energy. Let W_k contain the selected directions and $y = W_k^\top h^\ell(x)$ denote the projected activation.

In this subspace, we model the forget and retain activations using domain-specific means m_f and m_r with a shared regularized covariance Σ. Under equal domain priors, the Gaussian log odds are

$$\ell_{\mathrm{LDA},t}(x) = a^\top y + c, \qquad a = \Sigma^{-1}(m_f - m_r), \qquad c = -\frac{1}{2} a^\top (m_f + m_r). \tag{9}$$

In GATE-DU (O), the mature online gate combines the domain-LDA evidence with the preceding text–prototype gate:

$$\ell_{\mathrm{gate},t}(x) = (1 - \beta_t)\ell_{\mathrm{pre},t}(x) + \beta_t \ell_{\mathrm{LDA},t}(x). \tag{10}$$

The weight β_t is increased only when the proposed layer, discriminative subspace, boundary, and shadow decisions remain sufficiently stable across refreshes. The domain-LDA component is therefore treated as an optional visual refinement rather than an inevitable replacement for the text gate.

GATE-DU (V): retain-initialized visual gating. When labeled retain images are available from the outset, GATE-DU (V) applies the same S_B, S_W, layer-selection, and Gaussian-posterior construction directly to the labeled forget and retain support. The text and prototype warm-up stages are unnecessary, and the method begins with the domain-LDA Gaussian gate. GATE-DU (O) and

Table 1: OVDU performance across class fractions γ with a single forget domain. Results are averaged over the four forget-domain choices. F: forget accuracy (%, lower is better); R: macro retain-domain accuracy (%, higher is better); HM: harmonic mean (%, higher is better). Bold and underline denote the best and second-best results, respectively, in each column. All three proposed variants use steering strength $\lambda = 0.5$.

		$\gamma = 0.25$			$\gamma = 0.50$			$\gamma = 0.75$			$\gamma = 1.00$		
Dataset	**Method**	$F\downarrow$	$R\uparrow$	HM↑	$F\downarrow$	$R\uparrow$	HM↑	$F\downarrow$	$R\uparrow$	HM↑	$F\downarrow$	$R\uparrow$	HM↑
OfficeHome	NegGrad+Fisher	59.54	65.65	50.07	51.62	67.92	56.51	36.97	**71.90**	67.17	31.31	77.90	73.01
	ADU	21.97	27.61	40.79	28.26	49.22	58.38	33.63	67.96	67.16	36.37	**80.36**	71.02
	TMS	53.33	63.24	53.71	39.34	**68.33**	64.27	24.90	68.68	71.75	24.61	75.45	**75.42**
	GATE-DU(T)	**7.64**	36.97	52.80	**11.15**	40.78	55.90	**8.20**	36.02	51.74	**11.03**	40.99	56.12
	GATE-DU(O)	17.05	49.79	62.23	17.57	52.23	63.95	13.59	50.95	64.10	15.66	53.76	65.66
	GATE-DU(V)	27.64	**66.64**	**69.38**	26.66	66.14	**69.55**	23.37	67.86	**71.98**	19.76	70.10	74.83
DomainNet	NegGrad+Fisher	55.58	**77.72**	56.53	27.98	**77.08**	74.46	20.04	**77.24**	78.58	18.02	77.77	79.82
	ADU	14.64	24.45	38.01	19.68	49.25	61.06	23.41	67.27	71.63	27.86	**81.48**	76.53
	TMS	34.04	74.97	70.18	20.06	75.39	77.60	14.60	75.97	**80.41**	14.44	76.84	**80.97**
	GATE-DU(T)	**8.25**	62.50	**74.35**	**8.45**	62.55	74.32	**9.37**	64.38	75.28	**9.33**	63.82	74.92
	GATE-DU(O)	17.86	63.76	71.79	15.32	62.72	72.07	13.13	64.39	73.96	11.59	63.11	73.65
	GATE-DU(V)	22.10	70.60	74.07	16.39	73.39	**78.17**	14.92	74.59	79.49	12.98	75.09	80.62
PACS	NegGrad+Fisher	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
	TMS	76.08	90.39	37.82	47.06	93.49	67.60	47.07	93.24	67.53	38.35	**95.64**	74.97
	GATE-DU(T)	28.95	93.07	80.58	27.04	87.84	79.72	23.58	88.07	81.84	23.91	87.23	81.28
	GATE-DU(O)	**23.43**	92.84	**83.92**	21.87	75.09	76.58	17.78	78.02	80.06	25.78	93.19	82.63
	GATE-DU(V)	25.92	**93.08**	82.50	**21.21**	**93.89**	**85.68**	**16.62**	**94.20**	**88.46**	**14.84**	94.65	**89.65**

Table 2: Multi-domain OVDU performance on DomainNet when forgetting three domains ($|\mathcal{D}_f| = 3$). Results are averaged over the four possible forget-domain configurations. F: macro forget-domain accuracy (%, lower is better); R: retain-domain accuracy (%, higher is better); HM: harmonic mean (%, higher is better) respectively, in each column.

		$\gamma = 0.25$			$\gamma = 0.50$			$\gamma = 0.75$			$\gamma = 1.00$		
Dataset	**Method**	$F\downarrow$	$R\uparrow$	HM↑	$F\downarrow$	$R\uparrow$	HM↑	$F\downarrow$	$R\uparrow$	HM↑	$F\downarrow$	$R\uparrow$	HM↑
DomainNet	NegGrad+Fisher	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**
	TMS	7.11	60.19	73.05	5.07	62.45	75.34	4.81	66.10	**78.02**	4.59	67.21	78.86
	GATE-DU(V)	9.68	**61.94**	**73.49**	7.83	**66.43**	**77.21**	6.46	**66.83**	77.96	6.27	68.93	79.44

GATE-DU (V) therefore share the same mature visual discriminator but differ in how their retain statistics are obtained: GATE-DU (O) estimates them causally from pseudo-retain stream observations, whereas GATE-DU (V) constructs them directly from labeled retain support.

3.5 Shared Conditional Forgetting Payload

All three GATE-DU variants use the same class-agnostic forgetting payload. Once the corresponding gate evidence has been computed, forgetting is performed by contracting the selected activation toward the forget-domain centroid. Let $g_t(x) \in [0, 1]$ denote the gate activation and let $\lambda \in [0, 1]$ denote the steering strength. At payload layer ℓ, we apply

$$\widetilde{h}_t^{\ell}(x) = h^{\ell}(x) + \lambda g_t(x) \left(\mu_F^{\ell} - h^{\ell}(x)\right). \tag{11}$$

where $\mu_{F,\ell}$ is computed by pooling labeled forget support across classes. Since the target is class-agnostic, the payload does not require a separate steering direction for every class. Images assigned low forget evidence are left unchanged or only weakly modified, depending on the gate actuation used by the corresponding variant.

This shared construction makes the gate responsible for retain protection: the payload is not required to occupy a direction orthogonal to every retain-domain class feature. Open-vocabulary transfer is supported at three points in GATE-DU: the text gate marginalizes domain evidence over the class vocabulary, the visual discriminators use class-paired domain differences, and the payload uses a centroid pooled across the observed forget classes.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate our approach on three standard multi-domain benchmarks: OfficeHome (Venkateswara et al., 2017), comprising 65 classes across four domains; Mini-DomainNet (Peng et al., 2019), comprising 126 classes across four domains; and PACS (Li et al., 2017), comprising 7 classes across four domains. Following the few-shot protocol (Kawamura et al., 2026; Udupa et al., 2026), we use eight support images per observed class and domain for every dataset. The larger class vocabularies of OfficeHome and Mini-DomainNet enable a meaningful evaluation of open-vocabulary transfer from observed to held-out classes. In contrast, the small class inventory of PACS leaves very few observed classes at low vocabulary fractions; we therefore interpret it primarily as a low-sample stress test of the unlearning methods.

Baselines. We compare three variants of our conditional-steering framework: the retain-image-free text gate, its online visual-grounding extension, and the retain-initialized domain-LDA Gaussian gate. We benchmark these variants against the principal domain-unlearning baselines: Fisher-masked NegGrad+ (Udupa et al., 2026), the prompt-tuning-based ADU method (Kawamura et al., 2026), and TMS (Udupa et al., 2026), which introduces a preference-based objective for open-vocabulary domain unlearning. These baselines rely on gradient-based optimization and assume access to labeled retain-domain images. We preserve their original supervision setting and provide the same forget and retain support to our GATE-DU(V), enabling a supervision-matched comparison. By contrast, our GATE-DU(T) receives no retain image, while the GATE-DU(O) receives no labeled retain support and instead constructs visual retain evidence causally from confidently routed samples in the unlabeled evaluation stream.

Evaluation Protocol and Metrics. We follow the evaluation protocol introduced in (Udupa et al., 2026). Specifically, we adopt a strict few-shot setting with eight support images per observed class and domain, and evaluate class fractions $\gamma \in \{0.25, 0.50, 0.75, 1.00\}$. The setting with $\gamma = 1.00$ corresponds to closed-vocabulary domain unlearning, whereas $\gamma < 1.00$ evaluates open-vocabulary transfer at varying levels of class exposure. In every setting, evaluation is performed over the complete class vocabulary, including both classes observed during unlearning and held-out classes. We report forget-domain accuracy A_f and retain-domain accuracy A_r. To summarize their trade-off, we define the erasure score as $E = 100 - A_f$ and report the harmonic mean $H = 2EA_r/(E + A_r)$. All reported results are averaged across multiple independent runs to ensure statistical robustness. Additional details on the implementation on the exact entropy, readiness, refresh, and stability settings are provided in the Appendix A.2.

4.2 EXPERIMENTAL RESULTS

Overall efficacy of GATE-DU. Table 1 demonstrates that the training-free GATE-DU framework is both efficient and effective, particularly in open-vocabulary settings and the low-sample PACS regime. GATE-DU (V) matches or outperforms prior training-based DU methods. Despite requiring neither training nor retain images, GATE-DU (T) remains competitive on DomainNet and outperforms all prior methods on PACS. On OfficeHome, where text evidence is less informative, GATE-DU (O) improves upon GATE-DU (T).

Multi-domain forgetting. We evaluate multi-domain forgetting on DomainNet. Table 2 shows that GATE-DU (V) outperforms existing methods at low class fractions and remains competitive at higher fractions and in the closed-vocabulary setting, despite being training-free.

Sample efficiency. Fig. 1 compares GATE-DU(V) with prior DU methods using 1, 4, and 8 shots at class fractions 0.5 and 1.0 on Mini-DomainNet. Its performance largely stabilizes with a single shot, making it practical for real-world deployment. Additional experimental results are in the Appendix A.3.

Ablation study: When should GATE-DU (O) deploy online prototypes? Table 3 supports separating *visual-statistic collection* from *prototype deployment*. When the retain-image-free text gate is weak, the early prototype improves harmonic mean in all 12 OfficeHome Art, Product, and Real World runs. When the text gate is strong, the prototype hurts in all four OfficeHome Clipart runs and all 16 DomainNet runs. Replacing DomainNet's retain-domain admissions with oracle labels yields only a small improvement in prototype-only harmonic mean (0.649 to 0.658, versus 0.742 for frozen text at 50% class coverage), indicating that admission errors explain little of the gap. The on-

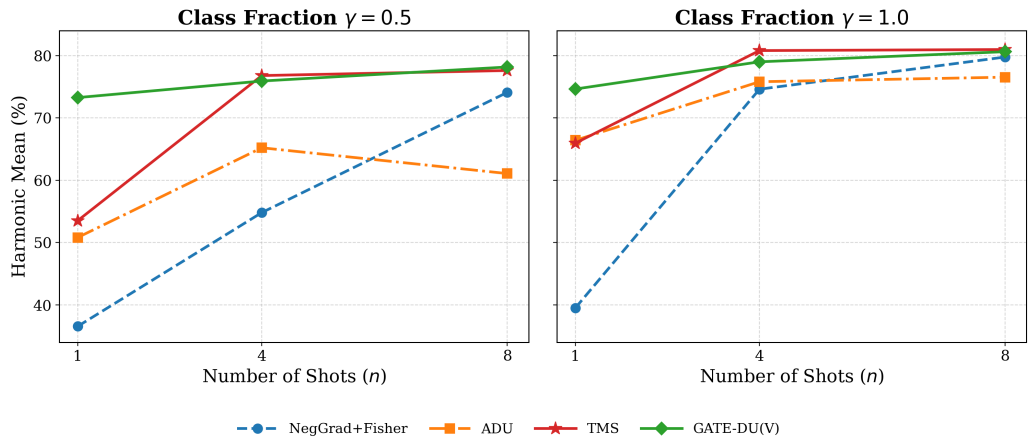


Figure 1: **Sensitivity to the Number of Training Samples.** Our training-free GATE-DU(V) method is stable using just a single shot of data sample while prior works needs at-least 4 shots and some require 8.

Table 3: The value of the prototype depends on the strength of the retain-image-free text gate. Results are averaged over four class fractions. Each ΔH is measured relative to frozen text (Harmonic Mean); Q4 denotes the final stream quartile.

Dataset	Forget domain	Text AUROC	ΔH_{proto}	ΔH_{full}	Q4 ΔH_{full}
OfficeHome	Art	0.687	+0.264	+0.265	+0.321
OfficeHome	Real World	0.712	+0.113	+0.113	+0.168
OfficeHome	Product	0.812	+0.067	+0.083	+0.117
OfficeHome	Clipart	0.942	−0.071	+0.023	+0.049
DomainNet	Clipart	0.922	−0.099	−0.030	+0.016
DomainNet	Painting	0.939	−0.061	−0.012	+0.036
DomainNet	Real	0.940	−0.100	−0.035	+0.012
DomainNet	Sketch	0.910	−0.099	+0.004	+0.057

line LDA stage substantially recovers this DomainNet loss: across all runs, harmonic mean increases from 0.656 with the prototype alone to 0.728 with LDA enabled, approaching the frozen-text value of 0.746. In the final stream quartile, the prototype-only branch remains 0.086 below frozen text, whereas the LDA-enabled pipeline is 0.031 above it. These results motivate a conditional policy: deploy the prototype early when independent evidence indicates a weak text gate; otherwise, retain the text-gate decisions while collecting visual statistics and activate LDA once a reliable candidate becomes available.

Why global subspace steering fails for domain unlearning? We compare domain unlearning with two class-unlearning controls. In **A**, the forget set contains all classes from Sketch, while the retain set contains the same classes from the other domains. In **B**, the 126 classes are divided into 63 forget and 63 retain classes, with both groups spanning all four domains. Thus, the domains overlap, but class identity separates the two groups. In **C**, the class split is retained, but forget samples come only from Sketch and retain samples only from Real, making both class and domain cues different. AlphaSteer is effective in the class-unlearning controls. Its best operating point achieves $F = 13.41$, $R = 73.17$, and $\text{HM} = 79.32$ in B, and $F = 10.48$, $R = 75.87$, and $\text{HM} = 82.14$ in C. Importantly, B remains effective despite sharing the same domain support, showing that a global subspace can capture class-discriminative structure. In contrast, the domain unlearning curves in Fig. 2 show forget and retain accuracy decreasing together under unconditional steering. Applying the same payload with an oracle domain gate reduces forget accuracy to 2.1% while preserving retain accuracy at 80.4%. These results suggest that CLIP admits useful global directions for class removal, but its domain representations are too entangled for a single unconditional subspace intervention. This motivates GATE-DU's separation of *when to intervene* from *how to steer*.

AI USE STATEMENT

In this work, we used generative AI tools for aiding in paper writing. Additionally, we used LLMs for literature survey and coding. To be precise, LLMs were used to check the related works in the field so that we do not miss any relevant work. We have not used generative AI tools for research formulation, experimental designs, analysis, generating synthetic datasets or for formulating or proving any scientific claims. We have reviewed all AI-assisted work. The coding agent's code was verified and tested for correctness. We take responsibility for the final content of this work including text produced with the aid of LLMs.

ETHICS STATEMENT

This work fully complies with the ICLR Code of Ethics. All research activities were performed with honesty, transparency, and a commitment to reproducibility. Experimental results are truthfully presented. No personal, sensitive, or proprietary data were used, and there are no privacy or licensing concerns associated with this study. No foreseeable harm arises from the work, prior literature and funding sources are cited, and there are no conflicts of interest. The paper does not infringe on institutional, legal, or ICLR ethical standards.

REPRODUCIBILITY STATEMENT

We provide complete details of the datasets, support construction, evaluation protocol, model configuration, hyperparameters, and online update procedure in the main paper and appendix. All hyperparameters are fixed across datasets unless explicitly identified as an ablation, and no evaluation labels are used for model selection or gate adaptation.

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)*, pp. 141–159. IEEE, 2021.

Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023.

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. In *International Conference on Learning Representations*, volume 2024, pp. 53643–53673, 2024.

Xinqi Fan, Xueli Chen, Luoxiao Yang, Chuin Hong Yap, Rizwan Qureshi, Qi Dou, Moi Hoon Yap, and Mubarak Shah. Test-time retrieval-augmented adaptation for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8810–8819. IEEE, 2025.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309. IEEE, 2020.

Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. Approximate domain unlearning for vision-language models. *Advances in Neural Information Processing Systems*, 38: 21805–21833, 2026.

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

Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research*, 3(4):6, 2024.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551. Ieee, 2017.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440. IEEE, 2024.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 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*, pp. 8748–8763. PmLR, 2021.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

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

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations*, volume 2026, pp. 25861–25904, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Sumanth Udupa, Mehrtash Harandi, Yadan Luo, and Mahsa Baktashmotlagh. Open vocabulary domain unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. To appear.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5385–5394. IEEE, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Barry Payne Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962.

Zixuan Weng, Jinghuai Zhang, Kunlin Cai, Ying Li, Peiran Wang, and Yuan Tian. Finesteer: A unified framework for fine-grained inference-time steering in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18736–18756, 2026.

Jing Wu and Mehrtash Harandi. Munba: Machine unlearning via nash bargaining. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–12. IEEE, 2025.

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

Xuanqi Zhang, Haoyang Shang, and Xiaoxiao Li. Gss: Gated subspace steering for selective memorization mitigation in llms. *arXiv preprint arXiv:2602.08901*, 2026.

A Appendix

A.1 Conclusion and Future Work

In this work, we propose GATE-DU, a unified training-free framework that formulates domain unlearning as conditional activation steering. Its three variants construct the gate from domain text, progressively ground it using an unlabeled stream, or initialize it directly from labeled retain images using domain-LDA and Gaussian calibration. Across open- and closed-vocabulary settings, our results show that deciding *when* to steer can provide an effective alternative to optimization-based unlearning while keeping the VLM and its prompts frozen.

Our current retain-image-free setting assumes that the names or descriptions of the retain domains are known. A natural future direction is fully open-domain unlearning, where neither retain images nor a complete retain-domain vocabulary is available. Developing gates that protect unknown and previously undescribed domains would make the task substantially harder, but also better reflect practical deployments in which the set of possible input domains cannot be enumerated beforehand.

A.2 Additional Implementation Details

A.2.1 Shared Experimental Setup

We implement all GATE-DU variants using a frozen CLIP ViT-B/16 model. Neither the vision encoder, text encoder, prompt embeddings, nor the forgetting payload is updated. Following the OVDU protocol, unlearning is performed using eight support images per selected class and domain. We vary the exposed class fraction as $\gamma \in \{0.25, 0.50, 0.75, 1.0\}$, while evaluation always includes the complete class vocabulary. Thus, $\gamma < 1$ measures transfer to classes that are not observed during gate construction.

For every vision block ℓ, let $H^\ell(x) \in \mathbb{R}^{T \times d_\ell}$ denote the residual-stream tokens, including the CLS token. Gate features are obtained by averaging over all tokens,

$$h^\ell(x) = \frac{1}{T} \sum_{i=1}^{T} H_i^\ell(x).$$

These activations are retained in their native space and are not ℓ_2-normalized. Direction vectors derived from them are normalized where required. The online prototype uses block 11, while the online LDA stage maintains statistics at blocks $\{9, 10, 11\}$. GATE-DU(V) searches the configured vision blocks and selects the block with the largest valid domain-LDA objective.

A.2.2 Forget-Support Partition

The selected forget classes are deterministically divided into two class-disjoint partitions, denoted by $\mathcal{C}_{\text{ref}}$ and $\mathcal{C}_{\text{cal}}$. The images belonging to $\mathcal{C}_{\text{cal}}$ form the calibration set $\mathcal{F}_{\text{cal}}$. This partition is used only to calibrate the initial text boundary; it does not remove these images from the remaining procedure.

After the text threshold and its scale have been fixed, all labeled forget-support images from both partitions are used to construct the class-wise forget statistics and the class-agnostic payload centroid. This separation prevents the same classes from defining and evaluating the initial text boundary while avoiding unnecessary loss of the limited visual support. For PACS at $\gamma = 0.25$, two classes are selected: one is used for text calibration and one remains class-disjoint, while both contribute to the subsequent visual statistics.

A.2.3 Calibration of GATE-DU(T)

Let $u_j = m_{\text{text}}(x_j)$ for $x_j \in \mathcal{F}_{\text{cal}}$. The text threshold τ_{text} is the largest threshold that accepts at least 90% of these forget-calibration samples. Operationally, if the u_j values are sorted in descending order, we use the $\lceil 0.90|\mathcal{F}_{\text{cal}}| \rceil$-th value. The resulting binary text gate is

$$g_{\text{text}}(x) = \mathbb{1}[m_{\text{text}}(x) \geq \tau_{\text{text}}].$$

The 90% target applies only to the support calibration set and does not force the evaluation stream to attain the same coverage.

For combination with the online visual scores, we convert the raw margin into a signed, dimensionless value:

$$\ell_{\text{text}}(x) = \text{clip}\left(\frac{m_{\text{text}}(x) - \tau_{\text{text}}}{s_{\text{text}}}, -12, 12\right).$$

The scale s_{text} is selected using the first finite value larger than 10^{-6} from

$$1.4826\ \text{MAD}(\{u_j\}), \qquad \frac{\text{IQR}(\{u_j\})}{1.349}, \qquad \text{Std}(\{u_j\}).$$

This operation does not fit a probability model. It only places the text margin on a stable scale for fusion with the prototype score. By construction,

$$\ell_{\text{text}}(x) \geq 0 \quad \Longleftrightarrow \quad m_{\text{text}}(x) \geq \tau_{\text{text}}.$$

GATE-DU(T) freezes this boundary and performs no stream adaptation.

A.2.4 Prequential Online Processing in GATE-DU(O)

GATE-DU(O) follows a strict test-then-update protocol. For stream item x_t, the complete prediction is produced using the immutable state S_{t-1}. Only after the gate decision and final prediction have been recorded may x_t update the statistics that define S_t. Therefore, an item can affect x_{t+1} and later samples, but never its own decision.

The processing order is:

1. Run the frozen model without steering and extract $z(x_t)$, the CLIP class distribution, and token-mean activations at blocks 9, 10, and 11.
2. Compute the frozen text evidence and any currently active prototype or LDA evidence using S_{t-1}.
3. Form the deployed gate score and record the gate decision, unsteered prediction, and final prediction.
4. Determine whether x_t is eligible for pseudo-retain admission.
5. If admitted, update the running statistics.
6. If a refresh condition is met, construct the next prototype or LDA candidate atomically.
7. Publish the updated state as S_t.

True stream-domain and class labels are accessed only after prediction for computing evaluation metrics. They never affect admission, candidate construction, reliability, or deployment.

A.2.5 Pseudo-Retain Admission

Let $q(c \mid x_t)$ be the frozen CLIP class distribution and

$$\hat{c}_t = \arg\max_c q(c \mid x_t)$$

its pseudo-class. Class confidence is measured using normalized entropy,

$$H_{\text{norm}}(x_t) = -\frac{\sum_c q(c \mid x_t) \log q(c \mid x_t)}{\log |\mathcal{C}|}.$$

An item is admitted as pseudo-retain only if

$$H_{\text{norm}}(x_t) \leq 0.35 \quad \text{and} \quad m_{\text{text}}(x_t) < \tau_{\text{text}}.$$

Thus, the text gate remains a fixed admission teacher even after the prototype or LDA becomes active. This avoids allowing an adaptive visual gate to generate its own pseudo-labels.

In the primary implementation, stream samples are used only as pseudo-retain evidence. The labeled forget support remains the trusted and fixed forget side. Samples failing either admission condition abstain and do not update any visual statistic.

A.2.6 STREAMING SUFFICIENT STATISTICS

For every pseudo-class c and tracked layer ℓ, we maintain a count $n^{\ell}_{r,c}$, mean $\mu^{\ell}_{r,c}$, and within-cell scatter $M^{\ell}_{2,r,c}$. Each admitted activation is incorporated using vector Welford updates (Welford, 1962). Given an activation h,

$$\delta = h - \mu, \qquad n^{+} = n + 1, \qquad \mu^{+} = \mu + \frac{\delta}{n^{+}},$$

$$M_2^{+} = M_2 + \delta(h - \mu^{+})^{\top}.$$

The statistics are accumulated in `float64`. We store no raw evaluation images, token tensors, or activation vectors after their corresponding update.

The labeled forget support initializes analogous per-class statistics $(n^{\ell}_{f,c}, \mu^{\ell}_{f,c}, M^{\ell}_{2,f,c})$, which remain fixed in the primary online experiment. A cell may maintain its count and mean after one observation, but it contributes to covariance estimation only after it contains at least two observations.

A.2.7 LOW-SAMPLE VISUAL PROTOTYPE

The prototype is constructed at block 11. A class is paired when both a trusted forget mean and a pseudo-retain mean are available:

$$\mathcal{P}_t = \left\{ c : n^{11}_{f,c} > 0 \ \wedge n^{11}_{r,c,t} > 0 \right\}.$$

A prototype candidate may be formed from one paired class, although the reliability terms prevent such a candidate from receiving substantial deployment weight.

Using the paired classes, we compute the shared direction

$$d_t = \frac{\sum_{c \in \mathcal{P}_t} \gamma_{c,t} \widehat{\Delta}_{c,t}}{\left\| \sum_{c \in \mathcal{P}_t} \gamma_{c,t} \widehat{\Delta}_{c,t} \right\|_2 + 10^{-12}}.$$

The boundary is the class-balanced mean of the projected forget–retain midpoints:

$$b_t = \sum_{c \in \mathcal{P}_t} \gamma_{c,t}\, d_t^{\top} \left(\frac{\mu^{11}_{f,c} + \mu^{11}_{r,c,t}}{2} \right).$$

This prevents frequently occurring pseudo-classes from dominating the intercept.

Let $S^{11}_{W,t}$ denote the pooled within-cell scatter reconstructed from the eligible forget and pseudo-retain cells. The prototype scale is

$$s_{\mathrm{proto},t} = \sqrt{\max\left(d_t^{\top} S^{11}_{W,t} d_t, 0 \right) + 10^{-6}}.$$

If this value is invalid, we use the standard deviation of the trusted forget projections and then a constant fallback of 1. The signed prototype evidence is

$$\ell_{\mathrm{proto},t}(x) = \mathrm{clip}\left(\frac{d_t^{\top} h^{11}(x) - b_t}{s_{\mathrm{proto},t}}, -12, 12 \right).$$

The orientation is fixed so that larger values are more forget-like.

Prototype statistics are updated after every admission, while a new candidate is constructed after every four newly admitted pseudo-retain samples. Candidate construction is atomic: an invalid candidate cannot partially update the direction, boundary, or scale.

A.2.8 TEXT-TO-PROTOTYPE TRANSITION

The prototype receives a deployment weight $\rho_t \in [0, 1]$. At prototype refresh r, we compute

$$R_{\mathrm{mass}} = \min\left(1, \frac{N_R}{8} \right), \qquad R_{\mathrm{pair}} = \min\left(1, \frac{|\mathcal{P}|}{4} \right),$$

where N_R is the total number of admitted pseudo-retain samples. Direction stability is measured using the signed cosine

$$S_{\text{dir}} = \text{clip}\left(d_r^\top d_{r-1}, 0, 1\right).$$

We do not square this value because opposite directions define opposite classifiers.

Boundary stability is computed as

$$z_b = \frac{b_r - b_{r-1}}{\max(s_{\text{proto},r}, s_{\text{proto},r-1}, 10^{-6})}, \qquad S_b = \exp\left(-\frac{z_b^2}{2}\right).$$

The first valid candidate receives zero direction and boundary stability because no preceding candidate exists. The target prototype weight is

$$\widehat{\rho}_r = \min\left(R_{\text{mass}}, R_{\text{pair}}, S_{\text{dir}}, S_b\right).$$

At refreshes, ρ moves 10% toward an equal or larger target and 50% toward a smaller target:

$$\rho \leftarrow \begin{cases} \rho + 0.10(\widehat{\rho} - \rho), & \widehat{\rho} \geq \rho, \\ \rho + 0.50(\widehat{\rho} - \rho), & \widehat{\rho} < \rho. \end{cases}$$

Between refreshes, ρ remains fixed. The pre-LDA score is

$$\ell_{\text{pre},t}(x) = (1 - \rho_t)\ell_{\text{text}}(x) + \rho_t \ell_{\text{proto},t}(x).$$

The prototype-only ablation uses this score with $\beta_t = 0$.

A.2.9 Online Domain-LDA Construction

The mature GATE-DU(O) variant reconstructs a domain-LDA gate from the same Welford statistics. A layer becomes eligible only when:

- both forget and pseudo-retain sides contain at least 32 samples;
- each paired domain–class cell contains at least two samples; and
- the number of paired classes is at least

$$\max(2, \min(8, |\mathcal{C}_{\text{support}}|)).$$

LDA reconstruction is attempted after every 32 new admissions once these conditions are met.

A.2.10 Online Gaussian Gate and Stability

Let W_k^ℓ contain the retained generalized eigenvectors and $y = W_k^{\ell\top} h^\ell(x)$. We compute projected forget and retain means m_f and m_r and a shared regularized covariance Σ. Equal domain priors are used irrespective of the number of admitted samples:

$$\pi_f = \pi_r = 0.5.$$

The Gaussian log odds are

$$\ell_{\text{LDA}}(x) = a^\top y + c, \qquad a = \Sigma^{-1}(m_f - m_r), \qquad c = -\frac{1}{2} a^\top (m_f + m_r).$$

In implementation, a is obtained through a linear solve; Σ^{-1} is never formed explicitly. Log odds are clipped to $[-30, 30]$ before evaluating the sigmoid.

We use four deterministic statistic shards to estimate candidate stability. For each leave-one-shard-out reconstruction, the generalized eigenvectors are converted to thin Euclidean QR bases only for comparing their spans. This gives a bounded subspace-consistency score based on

$$\frac{\|Q_i^\top Q_j\|_F^2}{k}.$$

The original generalized eigenvectors are retained for the Gaussian gate itself.

Each candidate is first evaluated in a shadow window without affecting predictions. At the next refresh, it can be committed only if the proposed layer persists and its direction, shadow decisions,

Table 4: Fixed implementation settings used by GATE-DU. These values are shared across datasets and are not selected using retain-evaluation performance.

Setting	Value
Backbone	Frozen CLIP ViT-B/16
Forget support	8 images per selected class
Class fractions	$\{0.25, 0.50, 0.75, 1.0\}$
Text calibration coverage	0.90
Text and prototype logit clipping	$[-12, 12]$
Prototype layer	11
Online LDA candidate layers	$\{9, 10, 11\}$
Pseudo-retain entropy threshold	0.35
Prototype refresh period	4 admissions
LDA refresh period	32 admissions
Prototype minimum paired classes	1
LDA samples per paired domain–class cell	2
LDA minimum samples per domain side	32
LDA minimum paired classes	$\max(2, \min(8, \lvert\mathcal{C}_{\text{support}}\rvert))$
LDA retained eigenvalue energy	90%
Stability shards	4
Reliability increase / decrease rates	$0.10/0.50$
Gaussian priors	$\pi_f = \pi_r = 0.5$
Gaussian log-odds clipping	$[-30, 30]$
Payload strength	$\lambda = 0.5$

posterior variation, and boundary remain sufficiently consistent. The target LDA weight is the minimum of these reliability terms. An invalid candidate retains the last valid gate and drives the target reliability toward zero.

The final online evidence is

$$\ell_{\text{gate},t}(x) = (1 - \beta_t)\ell_{\text{pre},t}(x) + \beta_t \ell_{\text{LDA},t}(x).$$

Like ρ, β is updated only at refreshes, moving 10% toward a larger target and 50% toward a smaller target. Hence an unstable LDA estimate is removed faster than it is introduced.

A.2.11 Retain-Initialized GATE-DU(V)

Unlike the hard text and online gates, this posterior continuously scales the payload. It introduces no target-coverage threshold or separately tuned sigmoid temperature: its midpoint and scale follow from the estimated domain means and shared covariance.

A.3 Additional Experiments

Scalability Test. To test the scalability of the GATE-DU(V), we implement our training-free framework on CLIP ViT-L/14 backbone and benchmark the performance on the Mini-Domainnet dataset against the relevant prior works. Tables 5 and 6 clearly showcase the efficacy of the proposed GATE-DU(V) method.
To test the scalability of GATE-DU(T), we test how the retain-image-free text gate scales with the performance of the models. Therefore, we test the scaling using three models with increasing order of performance (based on zero-shot domain classification accuracy): CLIP-ViT-B-16, CLIP-ViT-L-14, and MobileCLIP2-B (Faghri et al., 2025). Table 7 clearly follows the same order in unlearning performance as well giving near perfect correlation.

Additional Experiments on AlphaSteer. We provide the detailed plots and tables showcasing the effectiveness of AlphaSteer in exploiting the class discriminative subspaces but unable to work on the highly entangled domain subspaces in CLIP models.

While Fig. 2 quantitatively shows the failure of AlphaSteer in domain unlearning task, Fig. 3a and 3b showcase the qualitative reason behind it. The null channels that AlphaSteer selects to steer has very less selectivity (and therefore greater overlap between forget and retain subspaces).

Table 5: Comparison on DomainNet with ViT-L-14 ($K = 8$, $\gamma = 0.5$).

	Forget accuracy ↓			Retain accuracy ↑			Harmonic mean ↑		
Domain	Baseline	TMS	GATE-DU(V)	Baseline	TMS	GATE-DU(V)	Baseline	TMS	GATE-DU(V)
Painting	0.3016	0.2222	**0.1603**	**0.8603**	0.8254	0.8011	0.7710	0.8009	**0.8199**
Real	0.8651	0.2937	**0.1524**	**0.8386**	0.8069	0.7804	0.2324	0.7533	**0.8126**
Clipart	0.2714	**0.1508**	0.1651	**0.8354**	0.8164	0.7630	0.7784	**0.8325**	0.7973
Sketch	0.1635	**0.0841**	**0.0841**	**0.8228**	0.8074	0.8101	0.8296	0.8582	**0.8597**
Average	0.4004	0.1877	**0.1405**	**0.8393**	0.8140	0.7886	0.6529	0.8112	**0.8224**

Table 6: Comparison on DomainNet with ViT-L-14 ($K = 8$, $\gamma = 1.0$).

	Forget accuracy ↓			Retain accuracy ↑			Harmonic mean ↑		
Domain	Baseline	TMS	GATE-DU(V)	Baseline	TMS	GATE-DU(V)	Baseline	TMS	GATE-DU(V)
Painting	0.2222	0.1714	**0.1238**	**0.8566**	0.8381	0.8291	0.8153	0.8333	**0.8520**
Real	0.1460	0.1206	**0.0921**	0.8063	**0.8095**	0.7804	0.8295	**0.8430**	0.8394
Clipart	0.2079	0.1349	**0.1063**	**0.8360**	0.8190	0.7831	0.8134	**0.8414**	0.8347
Sketch	0.1222	**0.0746**	0.0825	**0.8349**	0.8212	0.8217	0.8558	**0.8702**	0.8669
Average	0.1746	0.1254	**0.1012**	**0.8335**	0.8219	0.8036	0.8285	0.8470	**0.8483**

Table 7: GATE-DU(T) on DomainNet with frozen text gating ($\gamma = 0.5$). F is forget-domain accuracy, $\bar{R}$ is mean accuracy over the retained domains, and H is the harmonic mean of $1 - F$ and $\bar{R}$. The Average row is over the four forget domains.

	ViT-B-16			ViT-L-14			MobileCLIP2-B		
Forget domain	$F \downarrow$	$\bar{R} \uparrow$	$H \uparrow$	$F \downarrow$	$\bar{R} \uparrow$	$H \uparrow$	$F \downarrow$	$\bar{R} \uparrow$	$H \uparrow$
Painting	**0.0190**	0.3931	0.5613	0.0540	0.6101	0.7418	0.0762	**0.6788**	**0.7826**
Real	**0.0476**	0.5683	0.7118	0.0698	**0.7011**	**0.7995**	0.0587	0.6767	0.7874
Clipart	0.0286	0.5249	0.6815	0.0254	0.5688	0.7183	**0.0238**	**0.6090**	**0.7501**
Sketch	**0.0397**	0.4296	0.5937	0.0698	0.6444	0.7614	0.0492	**0.8153**	**0.8779**
Average	**0.0337**	0.4790	0.6371	0.0548	0.6311	0.7553	0.0520	**0.6950**	**0.7995**

Table 8: AlphaSteer on the class-unlearning controls. In B, 63 forget and 63 retain classes both span Clipart, Painting, Real, and Sketch. In C, forget classes are drawn from Sketch and retain classes from Real, separating both class and domain. F: forget accuracy (%, lower is better); R: retain accuracy (%, higher is better); HM: forget–retain harmonic mean (%, higher is better). Bold and underline denote the best and second-best HM, respectively, within each experiment.

B: Class split, shared domains				**C: Class and domain split**			
λ	$F \downarrow$	$R \uparrow$	HM↑	λ	$F \downarrow$	$R \uparrow$	HM↑
0.00	80.56	80.40	31.32	0.00	73.02	88.57	41.37
0.50	46.43	81.75	64.73	0.25	40.63	84.13	69.61
0.60	34.13	80.63	72.51	0.30	27.94	82.22	76.81
0.75	20.95	77.54	78.29	0.35	17.46	80.63	81.58
0.80	17.94	75.79	78.80	0.40	10.48	75.87	**82.14**
0.85	15.32	74.44	79.23	0.45	8.25	72.06	80.72
0.90	13.41	73.17	**79.32**	0.50	4.76	63.17	75.96
0.95	11.11	71.59	79.31	0.60	0.32	47.30	64.16
1.00	9.60	69.29	78.45	0.75	0.63	18.73	31.52
				0.80	0.63	13.02	23.02
				0.85	0.63	8.57	15.78
				0.90	0.63	7.62	14.15
				0.95	0.63	5.08	9.66
				1.00	0.32	4.13	7.93

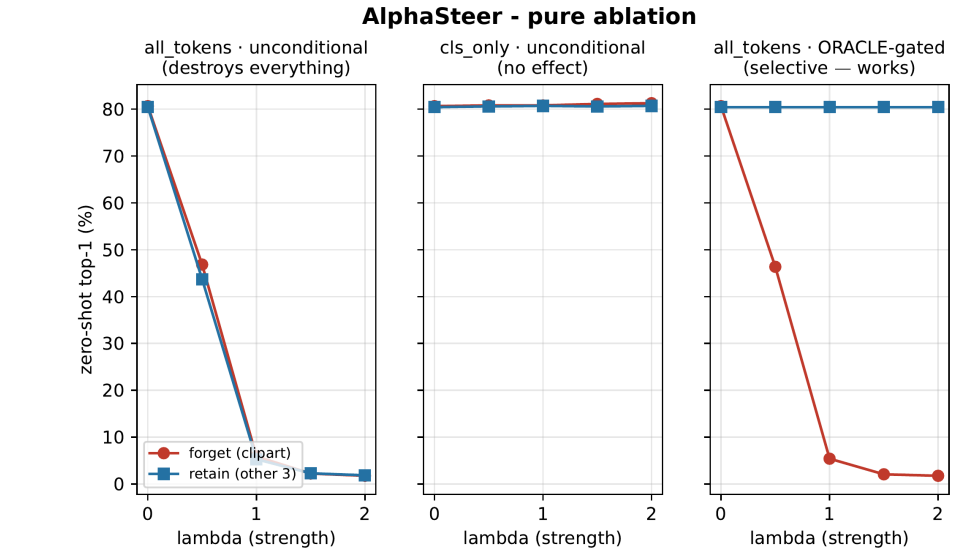
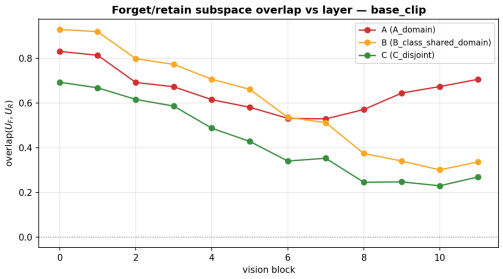
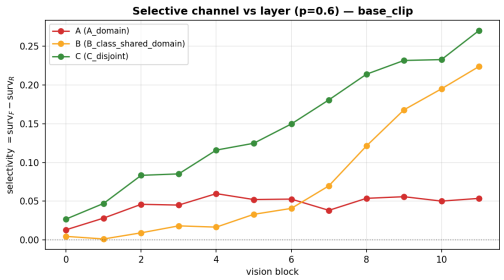


Figure 2: **Global AlphaSteer directions are insufficient for domain unlearning.** Applying the edit unconditionally to all tokens reduces forget and retain accuracy together (left), while editing only the CLS token has negligible effect (middle). Applying the same all-token payload only to forget-domain inputs using an oracle gate sharply reduces forget accuracy while preserving retention (right). This indicates that the payload is effective, but selectivity must primarily come from deciding *when* to steer.



(a) Layerwise overlap between the forget and retain domains in CLIP.



(b) AlphaSteer selectivity remains weak for domain unlearning (A) but increases for the class-unlearning controls (B and C), indicating that global subspaces capture class distinctions more readily than domain distinctions.

Figure 3: Qualitative reasoning as to why AlphaSteer struggles specifically with domain unlearning while retaining a viable operating point for class unlearning in VLMs.