TASK-AGNOSTIC DOMAIN UNLEARNING FOR VISION-LANGUAGE MODELS VIA CONTRASTIVE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) such as CLIP acquire domain-general image–text alignments, allowing the same semantic concept to be accessed across visual domains. This behavior becomes undesirable when a semantic association should be suppressed only in selected domains while remaining usable elsewhere. Existing domain unlearning approaches for CLIP are primarily evaluated through classification behavior and feature-level domain separation, which need not directly control the cross-modal ranking geometry used by both prompt-based classification and retrieval. We introduce Domain-Aware Contrastive Editing (DACE), a framework for approximate domain unlearning that directly edits domain-conditioned image–text relations. DACE preserves alignment on retain domains and applies one of three forgetting operators on forget domains: absolute similarity suppression (DACE-S), mismatch reassignment (DACE-M), or hard-negative margin reversal (DACE-H). We formalize the connection between alignment editing and downstream forgetting through cross-modal ranking margins: zero-shot classification and Recall@K are characterized by order statistics of the same similarity geometry, and the DACE-H hinge objective upper-bounds residual top-1 success on forget samples. This makes the edit task-independent at the level of the shared similarity geometry, with formal guarantees for prompt-based classification and top-K retrieval, without implying certified information removal. Across Office-Home, Mini DomainNet, and DomainNet, DACE suppresses forget-domain behavior while preserving retain-domain utility across classification and bidirectional retrieval with both CLIP and SigLIP. Representation and relearning analyses further indicate that behavioral forgetting can result from disrupting cross-modal accessibility without necessarily erasing all class-discriminative visual information.

1 INTRODUCTION

Vision-Language Models (VLMs) such as CLIP learn a shared embedding space aligning images and text, enabling strong cross-domain generalization and supporting a wide range of applications, including cross-modal retrieval, image–text generation, and multimodal reasoning (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; 2023; Liu et al., 2024; Karimian et al., 2025). As these models are increasingly deployed in downstream systems, retaining every learned association is not always desirable. Prior work shows that models may need to forget sensitive or proprietary information, outdated knowledge, or unsafe content, without sacrificing performance on the remaining data. Machine unlearning addresses this problem by removing the influence of designated data or concepts from a trained model while preserving utility on the remaining data (Cao & Yang, 2015; Bourtoule et al., 2021). Most existing unlearning methods for CLIP focus on class-level or concept-level forgetting (Kuwana et al., 2024). VLMs learn domain-general associations between images and text, allowing the same semantic concept to be recognized across real, artistic, sketch, or synthetic domains. While this cross-domain generalization is useful, it can be undesirable when recognition should be disabled only in specific domains. For example, a deployed system may be allowed to recognize real objects but required to suppress recognition or retrieval of synthetic, copyrighted, unsafe, or otherwise restricted visual domains. This setting differs from class-level unlearning: the concept itself should remain available, but its association with a designated domain should be weakened. We refer to this problem as approximate domain unlearning for VLMs.

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.

Recent work such as ADU (Kawamura et al., 2025) formulates approximate domain unlearning as reducing classification accuracy on forget domains while preserving performance on retain ones, highlighting the challenge of strong domain entanglement in pretrained VLM features. Existing methods define forgetting through classification objectives which do not explicitly control image–text alignment and may leave residual cross-modal associations. ADU (Kawamura et al., 2025) introduces a Domain Disentangling Loss (DDL) that encourages separation of domain-specific feature distributions in the latent space, addressing the strong entanglement of domains in pretrained VLM representations. However, this operates at the level of feature distributions rather than individual image–text pairs, and thus does not directly control pairwise image–text similarities. As a result, domain-level separation does not guarantee suppression of specific image–text correspondences.

In contrast to ADU (Kawamura et al., 2025), which formulates domain unlearning through classification objectives and feature-level domain disentanglement, we instead formulate domain unlearning in CLIP as a direct editing of cross-modal alignment in the shared embedding space. This yields a unified formulation where image–text alignment is preserved on retain domains and weakened on forget domains. Because prompt-based classification and cross-modal retrieval are both induced by relative image–text similarities, the edit is downstream-task-independent at the level of the shared similarity geometry rather than tied to a particular prediction head. We provide formal guarantees specifically for prompt-based classification and top-K retrieval. We do not claim task-agnosticity for arbitrary VLM capabilities such as generation or reasoning. We propose **DACE** (Domain-Aware Contrastive Editing), a family of methods for approximate domain unlearning that are designed to modify cross-modal alignment in CLIP's shared embedding space. On retain domains, we preserve the correct image–text pairing through a symmetric contrastive objective, while on forget domains, we modify this pairing through domain-conditional forgetting operators. We study three instantiations: **DACE-S**, which suppresses the similarity of the true image–text pair; **DACE-M**, which enforces reassignment to mismatched targets; and **DACE-H**, which imposes a margin-based constraint against the hardest negative. To localize the effect to target domains, we additionally incorporate a domain disentangling regularizer inspired by ADU. Experiments on multi-domain benchmarks demonstrate that this alignment-centric formulation yields strong domain unlearning performance and a favorable trade-off between retain-domain performance and forgetting effectiveness. We quantify this behavior using retain performance *Mem*, forgetting score *For*, and their harmonic mean *H*, where task-specific metrics (e.g., accuracy for classification and recall@K for retrieval) are used to instantiate these quantities. Our method improves this trade-off over prior approaches across both classification and cross-modal retrieval tasks within a unified framework.

Our main contributions are summarized as follows:

- We reformulate approximate domain unlearning in CLIP as *domain-conditional cross-modal relation editing*, where semantic associations are preserved in retain domains while their image–text accessibility is selectively disrupted in forget domains. We also introduce DACE, a unified family of alignment-editing objectives spanning absolute similarity suppression, mismatch reassignment, and hard-negative margin reversal. The formulation acts on the shared image–text geometry rather than on a downstream classifier.
- We formalize the link between alignment editing and behavioral forgetting through cross-modal ranking margins. We show that zero-shot classification and Recall@K are threshold events of order statistics of the same similarity geometry, and that the DACE-H hinge loss upper-bounds residual top-1 success on forget samples. This makes explicit why the same geometric edit transfers to these two downstream readouts, without claiming guarantees for arbitrary functions of the similarity ranking.
- Across classification and bidirectional retrieval, we show that a single edited model can suppress forget-domain behavior while preserving retain-domain utility. Representation and relearning analyses further distinguish disruption of cross-modal accessibility from complete erasure of visual semantic information.

2 Related Work

Machine Unlearning. Machine unlearning aims to remove the influence of designated data while preserving utility on retain data (Cao & Yang, 2015; Bourtoule et al., 2021; Sekhari et al., 2021). It is commonly divided into exact unlearning, which retrains after removing forget data (Guo et al.,

2020; Ginart et al., 2019; Neel et al., 2021), and approximate unlearning, which instead modifies a trained model using complementary forgetting and retain objectives (Golatkar et al., 2020; Thudi et al., 2022; Kurmanji et al., 2023; Tafti et al., 2026). Exact retraining is generally impractical for pretrained VLMs because of its computational cost and limited access to original training data (Kravets & Namboodiri, 2025a;b; Yang et al., 2025). Moreover, approximate unlearning methods are typically evaluated not only by their immediate post-unlearning performance but also by the persistence of forgetting, often through relearning attacks that test whether forgotten knowledge can be recovered (Ha et al., 2026; Dehghani et al., 2026).

Class Unlearning in CLIP. Early CLIP unlearning methods largely target class-level forgetting. Kravets & Namboodiri (2025a) extend local Lipschitz unlearning to CLIP using synthetic forget samples and selective layer updates, while Kravets & Namboodiri (2025b) weaken target-class associations through parameter-efficient adaptation of the text-side projection. Kuwana et al. (2024) enable black-box forgetting through derivative-free prompt optimization. These approaches improve practicality but mainly define forgetting through output-level behavior, such as reduced classification accuracy or increased uncertainty, which does not necessarily suppress the underlying image–text association in CLIP's contrastive space.

Cross-Modal Alignment Unlearning in CLIP. A complementary line of work targets CLIP's shared vision–language space directly. CLIPErase (Yang et al., 2025) removes designated visual–textual associations, Safe-CLIP (Poppi et al., 2024) suppresses unsafe concepts, RUCLIP (Zhang et al., 2025) improves robust concept unlearning through semantic anchors, and Han et al. (2025) remove noisy image–text correspondences. These methods demonstrate the importance of cross-modal alignment for unlearning, but primarily target concepts or associations globally. In contrast, our setting is domain conditional: the same semantic concept is preserved in retain domains while its cross-modal accessibility is selectively weakened in designated forget domains.

Domain Unlearning in CLIP. Kawamura et al. (2025) introduce approximate domain unlearning for CLIP, where recognition is suppressed in selected visual domains while preserved in others. Their ADU method combines cross-entropy minimization on retain domains, entropy maximization on forget domains, and a domain disentangling loss based on domain classification and MMD. This formulation addresses domain entanglement by separating domain distributions in the latent space, but forgetting is still defined through classification behavior and domain-level feature distributions rather than direct constraints on image–text correspondence. Consequently, reduced forget-domain classification accuracy does not necessarily imply that the underlying cross-modal alignment has been fully suppressed. This distinction is important because CLIP classification is a restricted ranking problem over text prompts, whereas image-to-text and text-to-image retrieval depend on the full image–text similarity matrix. We instead formulate domain unlearning as an instance-level modification of pairwise image–text alignment in CLIP's shared contrastive space, and evaluate forgetting through classification, cross-modal retrieval, and direct alignment changes.

3 METHODOLOGY

3.1 PROBLEM SETUP

We consider a multi-domain vision-language training set $\mathcal{S} = \{(\mathbf{x}_i, t_i, d_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ is an input image, $t_i \in \mathcal{T}$ is its associated text description, and $d_i \in \mathcal{D}$ is its domain label. We define retain-domain set $\mathcal{D}_r \subset \mathcal{D}$ as the set of domains to be preserved, and the forget-domain $\mathcal{D}_f = \mathcal{D} \backslash \mathcal{D}_r$ as the set of domains to be forgotten. Our goal is to adapt a pretrained CLIP model to preserve correct image–text alignment on samples $\{(\mathbf{x}, t, d) \mid d \in \mathcal{D}_r\}$, while weakening or breaking the image–text alignment on samples $\{(\mathbf{x}, t, d) \mid d \in \mathcal{D}_f\}$. We formulate domain unlearning in CLIP as a domain-conditional modification of cross-modal alignment. The objective is to preserve similarity between paired embeddings for samples in $\mathcal{D}_r$, while reducing or altering it for samples in $\mathcal{D}_f$. For a mini-batch $\mathcal{B} = \{(\mathbf{x}_i, t_i, d_i)\}_{i=1}^{|\mathcal{B}|}$, we denote by $\mathcal{B}_r$ and $\mathcal{B}_f$ the subsets of samples belonging to the retain and forget domains, respectively. Let f_v and f_t denote the image and text encoders of CLIP, which map inputs into a shared embedding space. For an input image $\mathbf{x}_i$, the normalized image representation is $z_i^V \in \mathbb{R}^m$, and for a text input t_i associated with $\mathbf{x}_i$, the normalized text representation is $z_i^T \in \mathbb{R}^m$. In the general setting, each image $\mathbf{x}_i$ is paired with a text description t_i, and learning is defined through alignment between (z_i^V, z_i^T) in the shared embedding space. When

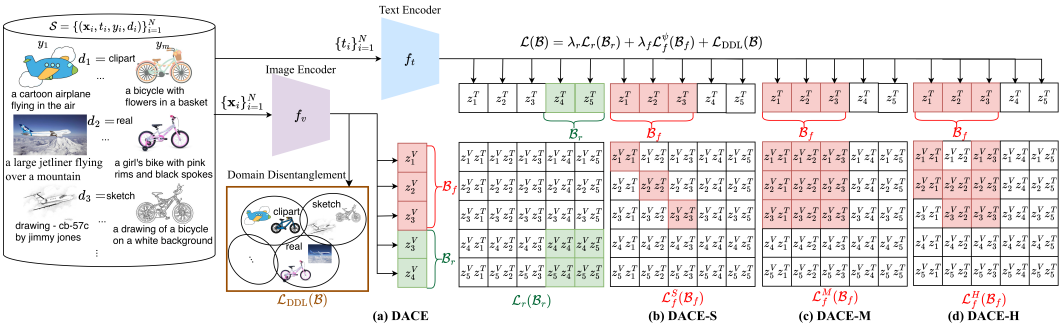


Figure 1: Overview of Domain-Aware Contrastive Editing (DACE), a family of methods for domain unlearning in CLIP, which preserves cross-modal alignment in retain domains $\mathcal{B}_r$ while weakening or breaking it in forget domains $\mathcal{B}_f$. (a) The full framework combines the retain loss $\mathcal{L}_r(\mathcal{B}_r)$, forget loss $\mathcal{L}_f(\mathcal{B}_f)$, and domain disentanglement loss $\mathcal{L}_{\text{DDL}}(\mathcal{B})$. (b) DACE-S weakens the similarity of true image-text pairs in $\mathcal{B}_f$. (c) DACE-M enforces stronger forgetting by reassigning samples in $\mathcal{B}_f$ to mismatched targets. (d) DACE-H performs targeted forgetting by ensuring that the similarity of the true image-text pair in $\mathcal{B}_f$ is lower than the hardest incorrect target by a margin m.

paired image–text supervision is unavailable, we use a class-prompt formulation, with the setup and results provided in Appendix B.

3.2 DACE: A Family of Task-Agnostic Domain Unlearning Methods

Our idea is to formulate domain unlearning in CLIP as a domain-conditional editing of image–text alignment in the contrastive space. Instead of defining forgetting through a task-specific classifier, we directly modify the visual-textual alignment. This yields **DACE** (**D**omain-**A**ware **C**ontrastive **E**diting), a family of domain-conditional alignment-editing objectives that share the same structure. As formalized in Appendix A, the formulation is downstream-task-agnostic for decision rules induced by the edited image–text similarity geometry. They differ in the choice of forgetting operator applied to the forget domains. The term *domain-aware* reflects that domain labels are used only during training to condition selective modification of image–text alignment on designated domains while preserving alignment on others, and are not used at inference. Formally, DACE is defined as

$$\mathcal{L}^{\text{DACE}}(\mathcal{B}) = \lambda_r \mathcal{L}_r(\mathcal{B}_r) + \lambda_f \mathcal{L}_f^{\psi}(\mathcal{B}_f) + \lambda_d \mathcal{L}_{\text{DDL}}(\mathcal{B}), \tag{1}$$

where λ_r and λ_f are weighting coefficients controlling the trade-off between retain and forgetting, while $\lambda_d \geq 0$ controls the optional domain-disentanglement regularization, and ψ parameterizes the specific forgetting operator. Here, $\mathcal{L}_r$ preserves correct image–text alignment on retain domains, $\mathcal{L}_f^{\psi}$ denotes the domain unlearning operator, and $\mathcal{L}_{\text{DDL}}$ is an optional domain-disentangling loss (DDL) shared across the family. For retain domains, all DACE variants preserve the correct cross-modal relation. To keep geometric margins distinct from temperature-scaled contrastive logits, define the cosine alignment score $a_{i,j} = z_i^{V\top} z_j^T$ and the corresponding contrastive logit $s_{i,j} = a_{i,j}/\tau$, where $\tau > 0$ is the temperature. With instance-specific paired captions, we use the standard symmetric paired contrastive objective

$$\mathcal{L}_r(\mathcal{B}_r) = \frac{1}{2}\left[\mathcal{L}_{v\to t}(\mathcal{B}_r) + \mathcal{L}_{t\to v}(\mathcal{B}_r)\right], \qquad \mathcal{L}_{v\to t}(\mathcal{B}_r) = -\frac{1}{|\mathcal{B}_r|}\sum_{i\in\mathcal{B}_r} \log \frac{\exp(s_{i,i})}{\sum_{j\in\mathcal{B}_r}\exp(s_{i,j})}, \tag{2}$$

and the text-to-image term defined analogously. Inspired by Kawamura et al. (2025), we use a DDL to encourage domain-separable latent structure:

$$\mathcal{L}_{\text{DDL}}(\mathcal{B}) = \alpha \mathcal{L}_{\text{CE}}(\mathcal{B}) - \beta \text{MMD}^2(\mathcal{B}), \tag{3}$$

where α and β control the strengths of the domain classification term ($\mathcal{L}_{\text{CE}}$) and pairwise inter-domain discrepancy term (MMD^2), respectively.

$$\mathcal{L}_{\text{CE}}(\mathcal{B}) = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{D}|} d_{i,j} \log p_{i,j}^d, \tag{4}$$

where $d_{i,j}$ is the one-hot domain label indicating whether sample i belongs to domain j, and $p^d_{i,j}$ denotes the predicted probability that sample i belongs to domain j. Let $\mathcal{D}_\mathcal{B} = \{d \in \mathcal{D} : \mathcal{B}_d \neq \emptyset\}$ denote the domains represented in the current mini-batch. For $|\mathcal{D}_\mathcal{B}| \geq 2$, the global discrepancy is

$$\mathrm{MMD}^2_{\mathrm{G}}(\mathcal{B}) = \frac{2}{|\mathcal{D}_\mathcal{B}|\big(|\mathcal{D}_\mathcal{B}|-1\big)} \sum_{\substack{d<d' \\ d,d' \in \mathcal{D}_\mathcal{B}}} \left\| \frac{1}{|\mathcal{B}_d|} \sum_{\mathbf{x}_i \in \mathcal{B}_d} \phi(u_i) - \frac{1}{|\mathcal{B}_{d'}|} \sum_{\mathbf{x}_j \in \mathcal{B}_{d'}} \phi(u_j) \right\|^2_\mathcal{H}. \tag{5}$$

Here, u_i denotes the feature representation of sample i, $\phi(\cdot)$ is a mapping to a reproducing kernel Hilbert space $\mathcal{H}$, and $\|\cdot\|_\mathcal{H}$ denotes the norm in that space. We refer to this represented-domain average as *global DDL* (G DDL). We also consider a retain–forget-aware variant (RF DDL), for which the discrepancy term is

$$\mathrm{MMD}^2_{\mathrm{RF}}(\mathcal{B}) = \left\| \frac{1}{|\mathcal{B}_r|} \sum_{i \in \mathcal{B}_r} \phi(u_i) - \frac{1}{|\mathcal{B}_f|} \sum_{j \in \mathcal{B}_f} \phi(u_j) \right\|^2_\mathcal{H}. \tag{6}$$

RF DDL replaces the pairwise all-domain discrepancy by the discrepancy between the aggregate empirical retain and forget feature distributions while retaining the domain-classification term. The negative sign in $\mathcal{L}_{\mathrm{DDL}} = \alpha\mathcal{L}_{\mathrm{CE}} - \beta\mathrm{MMD}^2$ is intentional: minimizing the objective simultaneously encourages domain predictability through $\mathcal{L}_{\mathrm{CE}}$ and increases the chosen inter-domain discrepancy. For global DDL, equation 5 averages only over domains represented in the current mini-batch; RF DDL is evaluated only for batches containing both retain and forget samples. The Reproducing Kernel Hilbert Space (RKHS) notation above leaves the particular positive-definite kernel and empirical estimator implementation-dependent; these must be kept identical to the implementation used to produce the reported experiments. Thus, all DACE variants share the same retain-domain alignment term $\mathcal{L}_r$ and the same domain-disentangling regularizer $\mathcal{L}_{\mathrm{DDL}}$, and differ only in the choice of forgetting operator $\mathcal{L}^\psi_f$. We next introduce three such variants.

3.3 DACE-S: Similarity Suppression

Our first instantiation, **DACE-S**, performs forgetting by directly suppressing the similarity of the true image–text pair on forget domains. Formally, for samples in $\mathcal{B}_f$, we define

$$\mathcal{L}^S_f(\mathcal{B}_f) = \frac{1}{|\mathcal{B}_f|} \sum_{i \in \mathcal{B}_f} \mathrm{sim}(z^V_i, z^T_i), \tag{7}$$

where $\mathrm{sim}(a,b) = a^\top b$ denotes the cosine similarity between ℓ_2-normalized embeddings. Minimizing this term weakens the correct positive-pair association on forget domains. The full DACE-S objective is therefore

$$\mathcal{L}^{\text{DACE-S}}(\mathcal{B}) = \lambda_r\mathcal{L}_r(\mathcal{B}_r) + \lambda_f\mathcal{L}^S_f(\mathcal{B}_f) + \lambda_d\mathcal{L}_{\mathrm{DDL}}(\mathcal{B}). \tag{8}$$

DACE-S can be viewed as the simplest member of the DACE family: it forgets by *suppressing* the original cross-modal alignment on forget domains. Because the embeddings are normalized, this linear objective has unconstrained minimum -1 and therefore, if optimized in isolation, would favor anti-alignment rather than merely zero similarity. In DACE it is optimized jointly with the retain objective and under a restricted adaptation parameterization; we therefore use "suppression" to describe its operational role rather than its isolated optimum. It should be interpreted as a minimal alignment-suppression operator. While it weakens the image–text association, it does not guarantee the complete removal of semantic information from the forget-domain sample. Instead, it primarily reduces the strength of the direct paired association used in CLIP-style classification and retrieval tasks. This limitation motivates the need for more aggressive variants, such as DACE-M and DACE-H, which enforce stronger constraints on relative alignment, thereby ensuring more effective forgetting of the target associations.

3.4 DACE-M: Mismatch Reassignment

Our second instantiation, **DACE-M**, breaks the original correspondence by redirecting forget-domain samples toward mismatched targets. For instance-specific paired captions, let π be a derangement of the forget mini-batch, i.e., a bijection $\pi : \mathcal{B}_f \to \mathcal{B}_f$ satisfying $\pi(i) \neq i$. This makes

the forward and reverse mismatched pairings well-defined. Using π for image-to-text matching and π^{-1} for the reverse direction, we define

$$\mathcal{L}^{\text{mis}}_{f,v\to t}(\mathcal{B}_f) = -\frac{1}{|\mathcal{B}_f|}\sum_{i\in\mathcal{B}_f}\log\frac{\exp(s_{i,\pi(i)})}{\sum_{j\in\mathcal{B}_f}\exp(s_{i,j})},\ \mathcal{L}^{\text{mis}}_{f,t\to v}(\mathcal{B}_f) = -\frac{1}{|\mathcal{B}_f|}\sum_{j\in\mathcal{B}_f}\log\frac{\exp(s_{\pi^{-1}(j),j})}{\sum_{i\in\mathcal{B}_f}\exp(s_{i,j})}. \tag{9}$$

Accordingly, the forget loss and thus, the full objective are as follows:

$$\mathcal{L}^{M}_{f}(\mathcal{B}_f) = \frac{1}{2}\left[\mathcal{L}^{\text{mis}}_{f,v\to t}(\mathcal{B}_f) + \mathcal{L}^{\text{mis}}_{f,t\to v}(\mathcal{B}_f)\right], \tag{10}$$

$$\mathcal{L}^{\text{DACE-M}}(\mathcal{B}) = \lambda_r\mathcal{L}_r(\mathcal{B}_r) + \lambda_f\mathcal{L}^{M}_{f}(\mathcal{B}_f) + \lambda_d\mathcal{L}_{\text{DDL}}(\mathcal{B}). \tag{11}$$

DACE-M performs target reassignment rather than suppression. Its effect depends on the chosen mismatch, motivating DACE-H, which adapts the competing target to the embedding geometry.

3.5 DACE-H: Hard-Negative Margin Forgetting

Our third instantiation, **DACE-H**, performs domain-conditional forgetting by enforcing that, on forget domains, the correct image–text alignment is no longer more similar than the most competitive incorrect alternative. Unlike DACE-S, which suppresses the true pair, and DACE-M, which enforces reassignment, DACE-H focuses the forgetting signal on the hardest negative under the current model. For each forget-domain sample $i \in \mathcal{B}_f$, let z^T_i denote its paired text embedding, which may correspond to a caption or a class-conditioned prompt depending on the available supervision. We define the hardest negative text for image i and the hardest negative image for text i as

$$h_t(i) = \arg\max_{j\in\mathcal{N}_t(i)}\text{sim}\big(z^V_i, z^T_j\big), \quad h_v(i) = \arg\max_{j\in\mathcal{N}_v(i)}\text{sim}\big(z^T_i, z^V_j\big) \tag{12}$$

where $\mathcal{N}_t(i)$ and $\mathcal{N}_v(i)$ contain only semantically incorrect candidates. With instance-specific captions, this excludes the paired index i; with class prompts, it excludes *all* candidates with class y_i, preventing same-class prompts from being treated as negatives. We define the image-to-text and text-to-image hard-negative margin losses as

$$\mathcal{L}^{H}_{f,v\to t}(\mathcal{B}_f) = \frac{1}{|\mathcal{B}_f|}\sum_{i\in\mathcal{B}_f}\max\Big(0,\ \text{sim}(z^V_i, z^T_i) - \text{sim}(z^V_i, z^T_{h_t(i)}) + m\Big), \tag{13}$$

$$\mathcal{L}^{H}_{f,t\to v}(\mathcal{B}_f) = \frac{1}{|\mathcal{B}_f|}\sum_{i\in\mathcal{B}_f}\max\Big(0,\ \text{sim}(z^T_i, z^V_i) - \text{sim}(z^T_i, z^V_{h_v(i)}) + m\Big), \tag{14}$$

where $m > 0$ is a margin hyperparameter. The full forgetting loss is given by

$$\mathcal{L}^{H}_{f}(\mathcal{B}_f) = \frac{1}{2}\left(\mathcal{L}^{H}_{f,v\to t}(\mathcal{B}_f) + \mathcal{L}^{H}_{f,t\to v}(\mathcal{B}_f)\right). \tag{15}$$

At zero loss, this objective enforces a *rank reversal*: an incorrect target must outrank the true target by at least m. Thus DACE-H controls a relative, decision-relevant margin rather than merely reducing an absolute similarity. Compared to DACE-M, which enforces explicit reassignment, DACE-H provides a more targeted and adaptive form of forgetting by focusing on the hardest negative under the current embedding geometry. The DACE-H objective is given by

$$\mathcal{L}^{\text{DACE-H}}(\mathcal{B}) = \lambda_r\mathcal{L}_r(\mathcal{B}_r) + \lambda_f\mathcal{L}^{H}_{f}(\mathcal{B}_f) + \lambda_d\mathcal{L}_{\text{DDL}}(\mathcal{B}). \tag{16}$$

When per-sample text representations are available, both image-to-text and text-to-image directions provide supervision and the objective is applied symmetrically.

Theoretical perspective. DACE-H edits a relative image–text ranking margin: at zero loss, an incorrect candidate in the training candidate set outranks the designated positive by at least the margin m. This connects the objective to prompt-based classification when all incorrect class prompts are included, and to retrieval ranking over the candidates against which the margin is evaluated. The complete mathematical treatment—including tie-safe classification and Recall@K characterizations, the top-K hinge-surrogate bound, and the sampled-negative-to-full-gallery certificate—is deferred to Appendix A. The theory concerns *alignment forgetting*: it certifies changes in ranking-based access to the designated positive, not deletion of all semantic information from the representation. Moreover, the sampled-negative full-gallery result requires an explicit negative-sampling assumption and is not used as an unconditional certificate for the reported retrieval experiments.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate DACE on three multi-domain benchmarks: Office-Home, with 65 classes across *art*, *clipart*, *product*, and *real* domains (Venkateswara et al., 2017); Mini DomainNet, with 126 classes across *clipart*, *painting*, *real*, and *sketch* domains (Zhou et al., 2021); and DomainNet, with 345 classes across *clipart*, *infograph*, *painting*, *quickdraw*, *real*, and *sketch* domains (Peng et al., 2019). We follow the standard eight-shot protocol with class and domain supervision, consistent with prior VLM adaptation and unlearning work (Zhou et al., 2022; Huang et al., 2024; Kawamura et al., 2025). For DomainNet retrieval, we use BLIP-2-generated captions from Kalluri et al. (2024). Further dataset and implementation details are provided in Appendix C.

Models. We use pretrained CLIP (Radford et al., 2021) with ViT-B/16 (Dosovitskiy et al., 2021). To assess generalization beyond CLIP and its contrastive objective, we extend DACE to SigLIP (Zhai et al., 2023), with the formulation, implementation details, and results reported in Appendix J.

Baselines. We compare our proposed DACE methods with zero-shot CLIP as the unedited reference, ADU (Kawamura et al., 2025)a prior approach to approximate domain unlearning in pretrained VLMs, and CLIPErase (Yang et al., 2025), a multimodal unlearning method that disrupts visual–textual associations in the forget set while preserving cross-modal alignment on retain set.

Evaluation metrics. Following prior work on approximate domain unlearning (Kawamura et al., 2025), we report retain accuracy on retain domains, $\mathit{Mem} = \mathrm{Acc}(\mathcal{D}_r)$, forgetting score as an error on forget domains, $\mathit{For} = 1 - \mathrm{Acc}(\mathcal{D}_f)$, and their harmonic mean $H = \frac{2\mathit{Mem}\cdot\mathit{For}}{\mathit{Mem}+\mathit{For}}$. For cross-modal retrieval tasks (text-to-image and image-to-text), we extend this formulation by replacing accuracy with retrieval metrics (e.g., R@K), and compute *Mem*, *For*, and *H* analogously.

4.2 RESULTS

In Table 1, we compare DACE with baselines across classification, text-to-image, and image-to-text retrieval. ADU achieves competitive classification harmonic means through strong forgetting; however, its retain retrieval performance is nearly zero, indicating a severe loss of cross-modal utility. A similar limitation is observed for CLIPErase: although its parameter updates effectively suppress forget image–text associations, its retain classification and retrieval performance deteriorates as the number of forget domains increases. Across the settings, DACE-H consistently achieves the strongest retrieval harmonic mean, demonstrating the most favorable balance between retain retrieval utility and forgetting among the parameter-updating approaches. When augmented with an additional class-prompt (CP) retain loss, DACE-H can also recover stronger classification performance while maintaining its retrieval advantage, as shown in Appendix F. DACE-S and DACE-M provide less aggressive alternatives, allowing the operating point to be adjusted toward greater retain performance or stronger forgetting. Together, these results support the alignment-margin view of Appendix A: editing relative image–text geometry can suppress similarity-induced behavior across multiple readouts while retaining cross-modal utility on preserved domains. Additional analyses include metric-wise retrieval results in Appendix D, per-domain results in Appendix M, representation and alignment analyses in Appendices H and I, and results on additional datasets in Appendix B.

4.3 ABLATION STUDIES

Effect of Domain Disentanglement Loss (DDL). Figure 2 ablates DDL on DomainNet with one forget domain ($|\mathcal{D}_f| = 1$). We compare three settings: **w/o DDL**, which removes the disentanglement regularizer; **G DDL**, which applies DDL globally to encourage stronger domain separation across the full training batch; and **RF DDL**, which applies DDL in a retain-forget-aware manner to better preserve retain-domain structure while separating forget-domain features. Results are reported for classification, text-to-image retrieval, and image-to-text retrieval in terms of *Mem*, *For*, and their harmonic mean (*H*). Overall, the impact of DDL is method-dependent. For DACE-S, adding DDL improves forgetting and generally leads to a better retain-forget trade-off, suggesting that disentanglement is particularly useful for the weaker suppression-based forgetting objective. In contrast, for DACE-M and DACE-H, the gains from DDL are smaller, since their alignment-editing objectives already impose a strong forgetting signal. Among the two disentanglement variants, RF

Table 1: Domain unlearning on DomainNet for classification and bidirectional retrieval, averaged over forget/retain domain configurations. Retrieval metrics average R@1, R@5, and R@10. **Bold** and underlined values denote the best and second-best results per metric within each $|\mathcal{D}_f|$ group.

$\lvert\mathcal{D}_f\rvert$	Method	Classification			Text-to-Image (Avg.)			Image-to-Text (Avg.)		
		Mem↑	*For*↑	*H*↑	*Mem*↑	*For*↑	*H*↑	*Mem*↑	*For*↑	*H*↑
1	Zero-shot CLIP	57.43	42.52	46.83	59.01	32.53	34.41	60.44	33.14	35.03
	ADU (Kawamura et al., 2025)	**59.96**	79.14	68.11	1.53	**98.95**	2.99	2.04	**99.17**	3.99
	CLIPErase (Yang et al., 2025)	25.86	**94.62**	40.27	29.21	93.97	43.32	30.83	93.99	44.91
	DACE-S (ours)	56.11	67.37	60.05	61.56	68.52	61.83	61.55	65.19	59.39
	DACE-M (ours)	54.67	90.29	68.06	58.89	85.60	68.19	60.50	89.36	70.72
	DACE-H (ours)	51.57	91.54	65.86	**78.84**	95.63	**85.20**	**78.57**	93.44	**84.48**
	DACE-H + CP retain (ours)	58.82	86.80	**70.11**	76.85	92.49	83.87	75.19	91.82	82.56
2	Zero-shot CLIP	57.41	42.67	48.40	60.66	35.17	39.51	61.86	34.69	39.41
	ADU (Kawamura et al., 2025)	**59.97**	81.22	68.98	1.74	**99.18**	3.41	2.74	**99.25**	5.28
	CLIPErase (Yang et al., 2025)	6.26	**98.76**	11.40	14.46	97.35	23.70	16.03	97.67	25.47
	DACE-S (ours)	55.17	65.16	59.66	61.74	68.11	62.50	61.52	67.00	61.32
	DACE-M (ours)	52.65	84.53	64.80	55.64	83.08	64.59	57.97	87.34	67.95
	DACE-H (ours)	41.64	91.30	56.89	**76.01**	95.23	**83.45**	73.75	94.46	82.12
	DACE-H + CP retain (ours)	58.38	85.79	**69.34**	75.58	92.17	82.84	**75.58**	91.86	**82.81**
3	Zero-shot CLIP	54.03	39.10	45.37	52.86	27.81	31.46	55.25	28.44	32.74
	ADU (Kawamura et al., 2025)	**57.44**	82.17	**67.61**	2.08	99.47	4.05	2.92	99.28	5.66
	CLIPErase (Yang et al., 2025)	0.31	**99.66**	0.63	0.15	**99.89**	0.31	0.25	**99.89**	0.50
	DACE-S (ours)	52.59	67.46	59.10	53.84	68.24	58.12	53.71	65.71	56.71
	DACE-M (ours)	49.06	82.70	61.59	45.78	83.40	56.78	49.39	87.24	61.18
	DACE-H (ours)	44.64	85.69	58.70	73.20	98.29	83.33	69.73	99.07	81.22
	DACE-H + CP retain (ours)	55.94	74.08	63.75	**74.20**	97.13	**84.13**	**72.53**	97.29	**83.10**
4	Zero-shot CLIP	**73.88**	50.91	60.28	74.28	42.47	50.47	78.22	43.20	52.34
	ADU (Kawamura et al., 2025)	68.73	85.67	76.27	2.98	99.50	5.76	4.40	99.62	8.38
	CLIPErase (Yang et al., 2025)	0.33	**99.69**	0.65	0.17	**99.94**	0.34	0.21	**99.93**	0.41
	DACE-S (ours)	68.98	67.47	68.21	73.98	66.42	67.35	73.37	73.87	71.47
	DACE-M (ours)	67.07	83.72	74.48	64.27	85.58	71.09	68.92	89.09	75.86
	DACE-H (ours)	61.88	86.80	72.25	**91.26**	97.78	**94.15**	**87.39**	98.94	**92.50**
	DACE-H + CP retain (ours)	68.26	88.69	**77.14**	84.25	95.14	89.37	84.59	95.67	89.79
5	Zero-shot CLIP	**66.09**	44.31	52.87	76.97	42.69	51.41	76.77	41.38	50.21
	ADU (Kawamura et al., 2025)	61.18	91.47	72.89	3.35	99.68	6.38	6.86	99.58	12.64
	CLIPErase (Yang et al., 2025)	0.33	**99.72**	0.66	0.17	**99.97**	0.34	0.21	**99.97**	0.41
	DACE-S (ours)	57.26	75.09	64.41	71.86	80.43	73.52	70.33	87.79	76.17
	DACE-M (ours)	59.57	89.04	70.89	66.75	89.75	74.45	69.12	95.00	78.36
	DACE-H (ours)	51.79	91.89	64.88	**89.70**	97.51	**93.13**	84.03	98.96	90.11
	DACE-H + CP retain (ours)	64.04	89.32	**74.40**	88.98	96.56	92.55	**85.31**	98.42	**91.30**

DDL gives the most consistent balance across tasks, whereas G DDL often increases forgetting more aggressively at the cost of lower memorization. Based on this ablation, we adopt RF DDL in the main experiments. The corresponding *Mem* and *For* results are reported separately in Appendix E.

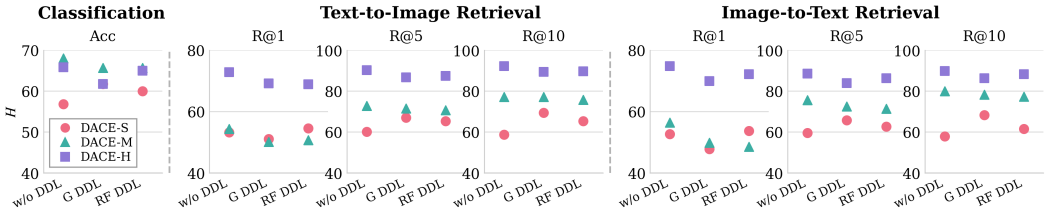


Figure 2: Ablation of DDL variants on the harmonic mean (*H*) on DomainNet with $|\mathcal{D}_f| = 1$, averaged over forget-domain choices. DACE-S, DACE-M, and DACE-H are evaluated without DDL, with G DDL, and with RF DDL across classification and bidirectional retrieval.

Contribution of individual loss components. Table 2 disentangles the roles of the retain, forgetting, and domain-disentanglement objectives by ablating $\mathcal{L}_r$, $\mathcal{L}_f$, and $\mathcal{L}_{\text{DDL}}$, using the RF DDL variant for $\mathcal{L}_{\text{DDL}}$. The results reveal complementary roles: $\mathcal{L}_f$ provides the primary forgetting signal, $\mathcal{L}_r$ is essential for preserving retain-domain alignment, and $\mathcal{L}_{\text{DDL}}$ serves as an auxiliary regularizer rather than the primary driver of forgetting. In particular, DACE-M and DACE-H remain effective without DDL, demonstrating that their forgetting behavior is primarily induced by the proposed alignment-editing objectives. For conciseness, retrieval results in Table 2 are averaged over

R@1, R@5, and R@10 and Appendix F reports metric-wise retrieval results and alternative retain classification losses that improve classification while preserving strong retrieval performance.

Table 2: Loss-component ablation on DomainNet for classification and bidirectional retrieval with $|\mathcal{D}_f| = 1$, averaged over forget/retain domain configurations. Retrieval metrics average R@1, R@5, and R@10. **Bold** and underlined values denote the best and second-best results per metric.

Method	$\mathcal{L}_r$	$\mathcal{L}_f$	$\mathcal{L}_{\text{DDL}}$	Classification			Text-to-Image (Avg.)			Image-to-Text (Avg.)		
				Mem↑	*For*↑	*H*↑	*Mem*↑	*For*↑	*H*↑	*Mem*↑	*For*↑	*H*↑
Retain-only DACE	✓	–	–	55.68	46.20	48.92	**84.59**	14.37	19.74	**84.10**	16.56	22.41
Forget-only DACE-S	–	✓	–	0.36	**99.76**	0.72	0.30	**99.96**	0.60	0.13	**99.82**	0.25
Forget-only DACE-M	–	✓	–	0.35	99.73	0.69	0.24	99.90	0.48	0.05	**99.82**	0.11
Forget-only DACE-H	–	✓	–	1.17	99.71	2.24	2.89	99.62	4.97	1.85	99.26	3.38
DACE-S w/o DDL	✓	✓	–	54.97	62.49	56.88	62.36	62.34	57.44	62.12	61.43	56.78
DACE-M w/o DDL	✓	✓	–	54.67	90.29	**68.06**	58.89	85.60	68.19	60.50	89.36	70.72
DACE-H w/o DDL	✓	✓	–	51.57	91.54	65.86	78.84	95.64	**85.20**	78.57	93.45	**84.48**
DACE-S	✓	✓	✓	**56.11**	67.37	60.05	61.56	68.52	61.83	61.55	65.19	59.39
DACE-M	✓	✓	✓	51.02	92.96	65.75	54.94	87.40	65.82	52.83	93.37	65.79
DACE-H	✓	✓	✓	50.03	93.38	65.01	74.88	94.23	82.11	74.94	93.59	82.34

Sensitivity to Number of Training Samples. We analyze the effect of training-set size on retain–forgetting trade-off. Figure 3 reports the harmonic mean (*H*) of *Mem* and *For* as a function of the number of shots across classification and cross-modal retrieval tasks. Across all tasks, we observe that increasing the number of shots consistently improves the trade-off for all DACE variants. In particular, DACE-H achieves the highest *H* in most settings, benefiting from stronger forgetting while maintaining competitive memorization, especially in retrieval tasks. DACE-M provides a balanced trade-off, while DACE-S remains more conservative, yielding slightly lower *H* due to weaker forgetting. Importantly, the relative ranking of the methods remains consistent across tasks and data regimes, indicating that the proposed alignment-based unlearning framework scales robustly with the amount of supervision. Additional results for *Mem* and *For* metrics are provided in Appendix G.



Figure 3: Harmonic mean (*H*) versus the number of shots on DomainNet with $|\mathcal{D}_f| = 1$, averaged over all possible choices of the forget domain, for classification and bidirectional retrieval.

5 Conclusion

We introduced Domain-Aware Contrastive Editing (DACE), a framework for approximate domain unlearning that edits domain-conditioned image–text relations in a VLM's shared embedding space. DACE spans absolute similarity suppression, mismatch reassignment, and hard-negative margin reversal. We characterized prompt-based classification and retrieval through cross-modal ranking margins, showing that they are different readouts of the same similarity geometry and that the DACE-H hinge objective bounds residual top-1 success on forget samples. Across Office-Home, Mini DomainNet, and DomainNet, DACE provides a better retain–forget trade-offs in classification and bidirectional retrieval. The representation and relearning analyses sharpen the interpretation of these results: behavioral forgetting can arise by disrupting cross-modal accessibility even when some class-discriminative visual information remains. Thus DACE should be viewed as approximate alignment unlearning rather than certified information erasure, while providing a mechanism for controlling domain-specific behavior in prompt-based classification and cross-modal retrieval through their shared ranking geometry. Limitations and future works are discussed in Appendix L.

AI USE STATEMENT

No generative AI tools were used in the preparation of this manuscript.

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.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Zahra Dehghani, Pablo Piantanida, and Mohammadhadi Shateri. Source-free class relearning: Diagnosing forgetting in class unlearning. *arXiv preprint arXiv:2609.02018*, 2026.

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

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems (NeurIPS)*, 32, 2019.

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 (CVPR)*, pp. 9304–9312, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, 13–18 Jul 2020.

SeungBum Ha, Saerom Park, and Sung Whan Yoon. Unlearning's blind spots: Over-unlearning and prototypical relearning attack. In *Forty-third International Conference on Machine Learning (ICML)*, 2026.

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

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23773–23782, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning (ICML)*, pp. 4904–4916. PMLR, 2021.

Tarun Kalluri, Bodhisattwa Prasad Majumder, and Manmohan Chandraker. Tell, don't show: Language guidance eases transfer across domains in images and videos. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Banafsheh Karimian, Giulia Avanzato, Soufian Belharbi, Alexis Guichemerre, Luke McCaffrey, Mohammadhadi Shateri, and Eric Granger. Clip-it: Clip-based pairing for histology images classification. *arXiv preprint arXiv:2504.16181*, 2025.

Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. Approximate domain unlearning for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 19113–19122, 2023.

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

Alexey Kravets and Vinay P Namboodiri. Zero-shot clip class forgetting via text-image space adaptation. *Transactions on Machine Learning Research (TMLR)*, 2025b.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems (NeurIPS)*, 36:1957–1987, 2023.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:28541–28564, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning (ICML)*, pp. 12888–12900. PMLR, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

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 (ICCV)*, pp. 1406–1415, 2019.

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision (ECCV)*, pp. 340–356. Springer, 2024.

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 (ICML)*, pp. 8748–8763. PmLR, 2021.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:18075–18086, 2021.

Zahra Dehghani Tafti, Pablo Piantanida, and Mohammadhadi Shateri. A universal source-free class unlearning framework via synthetic embeddings. *Transactions on Machine Learning Research (TMLR)*, 2026. ISSN 2835-8856.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

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 (CVPR)*, pp. 5018–5027, 2017.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 11975–11986, 2023.

Yue Zhang, Hanyu Li, Qinghong Yin, Xianlin Zhang, Ziyang Wang, and Xueming Li. Ruclip: Robust concept unlearning in clip via semantic anchors. *Expert Systems with Applications*, pp. 130495, 2025.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.

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

APPENDIX CONTENTS

A Alignment Margins and Task-Level Forgetting 13
B Classification-Specific Formulation and Additional Results 17
C Additional Implementation and Experimental Details 18
D Detailed Retrieval Results 19
E Additional DDL Ablation Results 21
F Ablation of Training Objectives 21
G Additional Ablation Results for Training Sample Sensitivity 23
H Representation-Level Analysis 24
I Visual Structure and Cross-Modal Alignment 26
J Generalization to SigLIP 27
K Robustness to Few-Shot Relearning Attacks 31
L Limitations and Future Work 32
M Per-Domain Results 32

A ALIGNMENT MARGINS AND TASK-LEVEL FORGETTING

The three DACE operators act on different aspects of the same cross-modal geometry. We now make explicit why relative alignment editing transfers to prompt-based classification and retrieval. Throughout this subsection, margins are defined using the *unscaled* cosine alignment

$$a_\theta(x,t) = \langle z_\theta^V(x), z_\theta^T(t) \rangle, \tag{17}$$

while the contrastive softmax uses the scaled logit $s_\theta(x,t) = a_\theta(x,t)/\tau$. This distinction avoids absorbing the temperature into the geometric margin.

Let t^+ be a designated correct textual target for image x, let $\mathcal{T}^-(x)$ contain semantically incorrect candidates, and define the top-1 alignment margin

$$\Delta_\theta(x,t^+) = a_\theta(x,t^+) - \max_{t^- \in \mathcal{T}^-(x)} a_\theta(x,t^-). \tag{18}$$

A positive margin places the correct target strictly above all incorrect candidates; a negative margin means that at least one incorrect target outranks it.

Definition 1 (Ranking-induced downstream rule)**.** *A downstream rule is* ranking-induced *if, for a candidate set $\mathcal{T} = \{t_1, \ldots, t_M\}$, its output depends on x only through the ordering of the similarity vector $A_\theta(x;\mathcal{T}) = (a_\theta(x,t_1), \ldots, a_\theta(x,t_M))$. Prompt-based zero-shot classification and top-K cross-modal retrieval are ranking-induced. The results below establish guarantees for these two readouts; they do not claim forgetting for every possible function of a similarity ranking. Accordingly, our formal use of* task-agnostic *means that the edit itself is performed at the level of the shared similarity geometry rather than a task-specific prediction head. It does not imply transfer to arbitrary VLM capabilities such as generation or reasoning.*

Proposition 1 (Classification as alignment ranking)**.** *For zero-shot classification $\hat{y}_\theta(x) = \arg\max_{c \in \mathcal{C}} a_\theta(x,t_c)$, define*

$$\Delta_\theta^{\mathrm{cls}}(x,y) = a_\theta(x,t_y) - \max_{c \neq y} a_\theta(x,t_c). \tag{19}$$

Whenever $\Delta_\theta^{\mathrm{cls}}(x,y) > 0$, the prediction is correct; whenever $\Delta_\theta^{\mathrm{cls}}(x,y) < 0$, it is incorrect. Hence, under any tie-breaking convention,

$$\Pr[\Delta_\theta^{\mathrm{cls}}(X,Y) > 0 \mid D_f] \leq \Pr[\hat{y}_\theta(X) = Y \mid D_f] \leq \Pr[\Delta_\theta^{\mathrm{cls}}(X,Y) \geq 0 \mid D_f], \tag{20}$$

with equality throughout when ties occur with probability zero.

Proof. A strictly positive margin means that the correct class has larger alignment than every competing class; a strictly negative margin means that at least one competing class has larger alignment. Only the zero-margin tie case depends on the tie-breaking rule, which gives equation 20. □

The implemented DACE-H loss uses a candidate set $\mathcal{N}_t(i)$. To state exactly what it controls, define the corresponding *training margin*

$$\Delta_{\theta,\mathcal{N}}(x_i, t_i) = a_\theta(x_i, t_i) - \max_{j \in \mathcal{N}_t(i)} a_\theta(x_i, t_j). \tag{21}$$

For class-prompt training, $\mathcal{N}_t(i)$ contains all incorrect class prompts, so $\Delta_{\theta,\mathcal{N}} = \Delta_\theta^{\mathrm{cls}}$. With mini-batch caption negatives, equation 21 certifies ranking only with respect to the candidates actually included in $\mathcal{N}_t(i)$.

Proposition 2 (DACE-H controls ranking within its training candidate set)**.** *For $m > 0$, let*

$$\ell_H(x, t^+) = [\Delta_{\theta,\mathcal{N}}(x, t^+) + m]_+. \tag{22}$$

Then

$$\mathbf{1}\{\Delta_{\theta,\mathcal{N}}(x, t^+) \geq 0\} \leq \frac{\ell_H(x, t^+)}{m}, \tag{23}$$

and therefore

$$\Pr_{D_f}[\Delta_{\theta,\mathcal{N}}(X, T^+) \geq 0] \leq \frac{\mathbb{E}_{D_f}[\ell_H(X, T^+)]}{m}. \tag{24}$$

In particular, $\ell_H(x, t^+) = 0$ implies $\Delta_{\theta,\mathcal{N}}(x, t^+) \leq -m$, so an incorrect candidate in $\mathcal{N}_t(x)$ outranks the correct target by at least m.

Proof. If $\Delta_{\theta,\mathcal{N}} < 0$, the indicator is zero. If $\Delta_{\theta,\mathcal{N}} \geq 0$, then $[\Delta_{\theta,\mathcal{N}} + m]_+ \geq m$. Dividing by m and taking expectations proves the claim. □

For class-prompt classification, where the training candidate set contains all incorrect class prompts, Propositions 1–2 imply

$$\mathrm{Acc}(D_f) \leq \frac{\mathbb{E}_{D_f}[\ell_H]}{m}, \qquad \mathrm{For}(D_f) \geq 1 - \frac{\mathbb{E}_{D_f}[\ell_H]}{m}, \tag{25}$$

where Acc and $\mathrm{For} = 1 - \mathrm{Acc}$ are normalized to $[0, 1]$; the tables report the corresponding percentages. This is a guarantee on ranking-induced behavior, not on removal of all information about the forgotten samples.

For retrieval, we first consider the single-designated-positive protocol used by the margin characterization below. Let $a^-_{\theta,(K)}(x)$ be the K-th largest alignment among incorrect targets and define

$$\Delta_\theta^{(K)}(x, t^+) = a_\theta(x, t^+) - a^-_{\theta,(K)}(x). \tag{26}$$

Proposition 3 (Recall@K as an order-statistic margin)**.** *For a query with one designated relevant target, if $\Delta_\theta^{(K)}(x, t^+) > 0$ then the correct target is in the top-K, while if $\Delta_\theta^{(K)}(x, t^+) < 0$ it is outside the top-K. Thus, under any tie-breaking convention,*

$$\Pr[\Delta_\theta^{(K)}(X, T^+) > 0] \leq \mathrm{R@K} \leq \Pr[\Delta_\theta^{(K)}(X, T^+) \geq 0], \tag{27}$$

with equality throughout when ties occur with probability zero.

Proof. A negative K-th order margin means that at least K incorrect targets have strictly larger alignment than the positive target, placing it outside the top-K. A positive margin means fewer than K incorrect targets can outrank it. The zero-margin case depends only on tie breaking. □

This characterization suggests the direct top-K surrogate

$$\ell_{H_K}(x, t^+) = [\Delta_\theta^{(K)}(x, t^+) + m]_+. \tag{28}$$

Theorem 1 (Margin-surrogate bound for ranking-based forgetting)**.** *For a query with one designated relevant target and any $m > 0$,*

$$\Pr[\Delta_\theta^{(K)}(X, T^+) > 0] \le \mathrm{R@K} \le \Pr[\Delta_\theta^{(K)}(X, T^+) \ge 0] \le \frac{\mathbb{E}[\ell_{H_K}(X, T^+)]}{m}. \tag{29}$$

Consequently, on the forget distribution,

$$\mathrm{For@K}(D_f) \ge 1 - \frac{\mathbb{E}_{D_f}[\ell_{H_K}(X, T^+)]}{m}, \tag{30}$$

where $\mathrm{For@K} = 1 - \mathrm{R@K}$ *is normalized to* $[0, 1]$.

Proof. The first two inequalities are Proposition 3. For the last one, if $\Delta_\theta^{(K)} \ge 0$ then $[\Delta_\theta^{(K)} + m]_+ \ge m$, whereas if $\Delta_\theta^{(K)} < 0$ the indicator of this event is zero. Hence $\mathbf{1}\{\Delta_\theta^{(K)} \ge 0\} \le \ell_{H_K}/m$ pointwise; taking expectations gives equation 29. The forgetting bound follows from $\mathrm{For@K} = 1 - \mathrm{R@K}$. □

A.1 From sampled negatives to a full retrieval gallery

The previous theorem assumes that the relevant order statistic is computed over the evaluation candidate set. In practice, DACE-H can instead be optimized using a smaller set of negatives, such as in-batch captions. The following result quantifies what can—and cannot—be inferred about a larger gallery from such sampled-negative training.

Fix a query (x, t^+) and a finite gallery of M semantically incorrect targets $\mathcal{G}^-(x) = \{t_1^-, \dots, t_M^-\}$. For a threshold $\eta \ge 0$, define the number of gallery negatives that outrank the positive by at least η,

$$R_\eta(x, t^+) = \sum_{j=1}^{M} \mathbf{1}\big\{a_\theta(x, t_j^-) \ge a_\theta(x, t^+) + \eta\big\}. \tag{31}$$

Let $\mathcal{N}_n$ be a subset of n negatives sampled uniformly without replacement from $\mathcal{G}^-(x)$ and let

$$\ell_{H,n}(x, t^+; \mathcal{N}_n) = \left[a_\theta(x, t^+) - \max_{t^- \in \mathcal{N}_n} a_\theta(x, t^-) + m\right]_+. \tag{32}$$

Theorem 2 (Sampled-negative certificate for full-gallery rank reversal)**.** *Fix $0 < \gamma < m$ and let*

$$\varepsilon_n(x, t^+) = \mathbb{E}_{\mathcal{N}_n}\left[\ell_{H,n}(x, t^+; \mathcal{N}_n)\right], \tag{33}$$

where the expectation is over uniform sampling without replacement. Then

$$\frac{\binom{M - R_{m-\gamma}(x,t^+)}{n}}{\binom{M}{n}} \le \min\left\{1, \frac{\varepsilon_n(x, t^+)}{\gamma}\right\}, \tag{34}$$

with the convention $\binom{q}{n} = 0$ for $q < n$. Consequently, for any $K \ge 1$, if

$$\frac{\varepsilon_n(x, t^+)}{\gamma} < \frac{\binom{M-K+1}{n}}{\binom{M}{n}}, \tag{35}$$

then $R_{m-\gamma}(x, t^+) \ge K$. Hence at least K gallery negatives outrank the positive by the strictly positive margin $m - \gamma$, and the positive target is outside the full-gallery top-K.

Proof. Let E_γ be the event that the sampled set contains no negative with alignment at least $a_\theta(x, t^+) + m - \gamma$. On E_γ,

$$\max_{t^- \in \mathcal{N}_n} a_\theta(x, t^-) < a_\theta(x, t^+) + m - \gamma,$$

up to a boundary equality that can be included without changing the argument by using the corresponding weak event. Therefore $\ell_{H,n} \ge \gamma$ on E_γ, and Markov's elementary reverse implication gives

$$\Pr(E_\gamma) \le \frac{\varepsilon_n(x, t^+)}{\gamma}. \tag{36}$$

Exactly $R_{m-\gamma}$ of the M gallery negatives meet the threshold. A uniformly sampled subset of size n misses all of them with hypergeometric probability

$$\Pr(E_\gamma) = \frac{\binom{M-R_{m-\gamma}}{n}}{\binom{M}{n}},$$

which proves equation 34. If $R_{m-\gamma} \le K-1$, monotonicity of the binomial coefficient in its upper argument gives

$$\Pr(E_\gamma) \ge \frac{\binom{M-K+1}{n}}{\binom{M}{n}}.$$

Thus equation 35 contradicts $R_{m-\gamma} \le K-1$, so $R_{m-\gamma} \ge K$. Since $m-\gamma > 0$, these K negatives strictly outrank the positive, placing it outside the top-K. □

Theorem 2 makes the sampling gap explicit. A single zero-loss minibatch certifies only that at least one sampled negative outranks the positive by margin m; it does not certify a full-gallery top-K failure for $K > 1$. In contrast, a small *expected* sampled-negative hinge over repeated random candidate sets forces the gallery to contain many sufficiently strong negatives. The dependence on n is also explicit: larger negative sets make it easier to detect the gallery's high-similarity tail. This conclusion requires the sampling model in the theorem. In-batch negatives need not be exactly uniform over the evaluation gallery, so we use the result as a bridge clarifying the role of negative coverage rather than as an unconditional certificate for our reported retrieval numbers.

A.2 Distributional form

An analogous statement holds when negatives are sampled i.i.d. from a proposal distribution $Q(\cdot \mid x)$. Define

$$q_\eta(x, t^+) = \Pr_{T^- \sim Q(\cdot|x)}\left[a_\theta(x, T^-) \ge a_\theta(x, t^+) + \eta\right]. \tag{37}$$

For n independent negatives, the probability that none crosses the $(m-\gamma)$ threshold is $(1-q_{m-\gamma})^n$. Repeating the preceding argument yields

$$q_{m-\gamma}(x, t^+) \ge 1 - \left(\min\left\{1, \frac{\varepsilon_n(x, t^+)}{\gamma}\right\}\right)^{1/n}. \tag{38}$$

Thus sampled DACE-H controls a lower bound on the mass of negatives that outrank the forgotten positive by a positive margin. If an evaluation gallery of M negatives is itself sampled i.i.d. from the same Q, then, conditional on (x, t^+), the number of such negatives is binomial. Writing the right-hand side of equation 38 as $\underline{q}_\gamma(x, t^+)$ gives

$$\Pr[t^+ \text{ is in top-}K \mid x, t^+] \le \sum_{r=0}^{K-1} \binom{M}{r} \underline{q}_\gamma^{\,r} (1-\underline{q}_\gamma)^{M-r}. \tag{39}$$

The bound is intentionally conditional and assumption-explicit: it connects sampled-negative optimization to full-gallery ranking only when training negatives and gallery negatives follow the same proposal law. Distribution shift or biased in-batch sampling can invalidate this transfer.

Our experiments use the original DACE-H, corresponding to a $K = 1$ training margin over its candidate set. Importantly, minimizing this $K = 1$ objective does *not* by itself imply suppression of R@5 or R@10; the strong R@5/R@10 behavior observed experimentally is therefore empirical rather than a consequence of Theorem 1. equation 28 is a theoretically motivated extension that would directly target a chosen K, not an additional experimentally validated variant. If a benchmark admits multiple relevant targets per query, the corresponding event should instead be defined using the highest-ranked relevant target; the single-positive statement above is not silently assumed for such protocols.

A.3 PROBABILISTIC INTERPRETATION

Under the contrastive softmax pairing weights

$$p_\theta(t_j \mid x) = \frac{\exp(a_\theta(x, t_j)/\tau)}{\sum_k \exp(a_\theta(x, t_k)/\tau)}, \tag{40}$$

zero DACE-H loss against hardest negative t_h implies

$$\frac{p_\theta(t^+ \mid x)}{p_\theta(t_h \mid x)} = \exp\left(\frac{a_\theta(x, t^+) - a_\theta(x, t_h)}{\tau}\right) \leq \exp(-m/\tau). \tag{41}$$

Thus m has a direct contrastive-odds interpretation: increasing the unscaled cosine margin exponentially suppresses the softmax weight assigned to the original positive relative to its strongest incorrect competitor. This is a property of the contrastive normalization over the chosen candidate set, not a claim that these weights are calibrated semantic-match probabilities.

A.4 WHAT IS AND IS NOT FORGOTTEN

The margin theory characterizes *alignment forgetting*: DACE edits whether a semantic target remains accessible through the image–text ranking geometry. It does not imply that all information about class or instance identity has been removed from the visual representation. This distinction motivates the representation and relearning analyses in Appendix H, which test whether class-discriminative visual structure can remain even after cross-modal access has been disrupted.

B CLASSIFICATION-SPECIFIC FORMULATION AND ADDITIONAL RESULTS

In classification settings where paired text descriptions are not available and only class labels $y_i \in \mathcal{C}$ are given, we associate each sample $(\mathbf{x}_i, y_i)$ with a text prompt t_{y_i}. Specifically, for each class $y \in \mathcal{C}$, we define a class-dependent text prompt t_y and compute its normalized embedding z_y^T. In this case, for each training sample $(\mathbf{x}_i, y_i)$, we define a per-sample text target as $\tilde{z}_i^T = z_{y_i}^T$. Under this class-prompt formulation, the general DACE objectives are instantiated as follows.

DACE retain loss: When supervision is provided by class prompts, several images in a batch can share the same text target. Treating those duplicates as negatives would create false negatives. We therefore use unique class prompts and the class-conditional image-to-text loss

$$\mathcal{L}_r^{\mathrm{CP}}(\mathcal{B}_r) = -\frac{1}{|\mathcal{B}_r|} \sum_{i \in \mathcal{B}_r} \log \frac{\exp(s_{i,y_i})}{\sum_{c \in \mathcal{C}} \exp(s_{i,c})}, \tag{42}$$

where $a_{i,c} = z_i^{V\top} z_c^T$ and $s_{i,c} = a_{i,c}/\tau$. The denominator ranges over the full dataset class set $\mathcal{C}$, not merely classes represented in the mini-batch. In this setting the text-to-image direction is omitted because a class prompt corresponds to multiple images rather than a unique paired instance. This convention is used throughout the class-prompt formulations below.

DACE-M forget loss: For class-prompt supervision, a sample-level derangement with a different-class constraint need not exist (e.g., for class-imbalanced mini-batches). We therefore define the mismatch directly at the class level. For each sample i, choose a target class $\widetilde{y}_i \in \mathcal{C} \setminus \{y_i\}$ (deterministically or from a distribution supported on incorrect classes) and use

$$\mathcal{L}_{f,\mathrm{cls}}^M(\mathcal{B}_f) = -\frac{1}{|\mathcal{B}_f|} \sum_{i \in \mathcal{B}_f} \log \frac{\exp(s_{i,\widetilde{y}_i})}{\sum_{c \in \mathcal{C}} \exp(s_{i,c})}. \tag{43}$$

This construction is always well-defined whenever $|\mathcal{C}| \geq 2$ and guarantees that a same-class prompt is never used as the mismatch. In class-prompt experiments, $\mathcal{L}_f^M$ denotes equation 43.

DACE-H forget loss: In classification-only settings, where text representations are instantiated from fixed class prompts rather than instance-specific paired texts, we use only the image-to-text component $\mathcal{L}_{f,\mathrm{cls}}^H = \mathcal{L}_{f,v\to t}^H$, since the reverse direction does not provide additional supervision. So, in class-prompt settings, the forgetting objective is applied in the image-to-text direction. We

next formalize the class of downstream tasks for which editing this similarity geometry transfers directly to behavioral forgetting.

Together, these class-prompt formulations define the classification instantiation of DACE used in our experiments. We next evaluate their domain unlearning behavior across the classification benchmarks. Table 3 reports classification results for domain unlearning. Across Office-Home and Mini DomainNet, DACE improves the retain–forget trade-off over zero-shot CLIP and ADU in most settings. In particular, DACE-S achieves the strongest retain performance while still providing effective forgetting on the designated domains. By contrast, DACE-M and DACE-H impose stronger forgetting, typically at the cost of some retain accuracy, but often yield a better harmonic mean (H) because of this more aggressive unlearning behavior. Overall, these results show that DACE provides a flexible balance between preserving retain-domain performance and suppressing unwanted domain-specific associations. The corresponding per-domain results for each forget-domain configuration are reported in Appendix M.

Table 3: Comparison of domain unlearning methods on Office-Home and MiniDomainNet, reporting performance aggregated over all possible combinations of forget and retain domains for different numbers of domains to forget, $|\mathcal{D}_f| \in \{1, 2, 3\}$, as mean. Within each dataset, $|\mathcal{D}_f|$ group, and metric column, **bold** indicates the best mean result and underlined indicates the second-best mean result.

Dataset	Method	$\lvert\mathcal{D}_f\rvert = 1$			$\lvert\mathcal{D}_f\rvert = 2$			$\lvert\mathcal{D}_f\rvert = 3$		
		Mem↑	*For*↑	*H*↑	*Mem*↑	*For*↑	*H*↑	*Mem*↑	*For*↑	*H*↑
Office-Home	Zero-shot CLIP	82.54	17.45	28.07	**82.53**	17.47	28.58	**82.55**	17.46	28.79
	ADU (Kawamura et al., 2025)	80.08	61.78	69.18	78.26	68.44	72.82	73.62	77.69	74.92
	DACE-S (ours)	**82.64**	67.96	73.18	79.08	73.44	75.69	77.19	83.41	79.93
	DACE-M (ours)	76.60	78.34	77.22	72.87	80.29	75.90	70.49	84.03	76.28
	DACE-H (ours)	78.16	**87.74**	**82.62**	74.86	**88.59**	**81.02**	73.77	**88.08**	**80.04**
MiniDomainNet	Zero-shot CLIP	**82.78**	17.22	28.23	**82.78**	17.22	28.45	**82.78**	17.22	28.51
	ADU (Kawamura et al., 2025)	81.27	72.49	76.57	78.93	77.06	77.93	77.33	83.21	79.98
	DACE-S (ours)	81.94	85.26	83.53	79.43	87.67	**83.30**	79.17	91.37	**84.74**
	DACE-M (ours)	81.19	87.45	**84.17**	78.16	86.97	82.28	76.92	87.99	82.00
	DACE-H (ours)	77.95	**91.49**	84.09	75.78	**92.20**	83.07	68.74	**94.21**	78.43

C Additional Implementation and Experimental Details

During training, DACE requires an image and an associated text target for each sample. When image–text pairs are unavailable, as in Office-Home and Mini DomainNet, we instantiate the text target from the class label using the standard CLIP prompt template "a photo of a [class]". For DomainNet, where paired captions are available, we use the caption associated with each image as the training text target. This ensures that the training objective remains unchanged across datasets: DACE consistently operates by editing the alignment between each image and its corresponding text representation, while the source of the text target depends on the available annotations. For evaluation, we adapt the task through the choice of text queries. On Office-Home and Mini DomainNet, we evaluate classification only, since these datasets do not provide image–text pairs. Classification follows the standard zero-shot CLIP protocol, using "a photo of a [class]" as the text query for each class. On DomainNet, we evaluate classification as well as text-to-image and image-to-text retrieval. Classification again uses class-prompt queries, while retrieval is evaluated using paired image–text data by ranking candidates based on image–text similarity. For all proposed DACE variants, we set the retain and forget weights in equation 1 to $\lambda_r = 1$ and $\lambda_f = 1$. For the domain disentanglement term, we consider two settings. In the retain-forget-aware variant (RF DDL), which encourages separation between retain and forget domains, we use a domain-classification weight of $\alpha = 10$ and an MMD weight of $\beta = 1$. In the global variant (G DDL), which enforces separation across all domains, we set $\alpha = 1$ and $\beta = 1$. For DACE-H, the hard-negative margin in equation 13 and equation 14 is set to $m = 0.1$. Unless otherwise noted, all main results for the proposed methods are reported using RF DDL. For all experiments, the pretrained CLIP image and text encoders are frozen, and only lightweight trainable modules are updated. In the classification-only setting, we update the visual prompts, the learnable text prompt context, the domain classifier used for domain disentanglement, and the inserted attention/projection modules. In the task-agnostic setting, which is evaluated on classification, text-to-image retrieval, and image-to-text retrieval, we disable text prompt learning by setting the number and depth of text prompts to zero. The paired text/caption embeddings are therefore computed with the frozen CLIP text encoder, while the visual prompts,

domain classifier, and inserted attention/projection modules remain trainable. Training for Table 3 is performed for 150 epochs, while for Table 1, training is performed for 50 epochs, both using SGD with a learning rate of 0.0025. On the vision side, we employ deep visual prompting (Khattak et al., 2023) with eight learnable context tokens. The visual prompts are applied to and optimized within the first nine transformer layers of the image encoder.

For Office-Home and Mini DomainNet, we evaluate all possible forget-domain combinations, yielding 4, 6, and 4 configurations for $|\mathcal{D}_f| = 1, 2$, and 3, respectively. For DomainNet, we evaluate all configurations when $|\mathcal{D}_f| = 1$, and use a subset of configurations for larger forget sets: 3 for $|\mathcal{D}_f| = 2$, 1 for $|\mathcal{D}_f| = 3$, 1 for $|\mathcal{D}_f| = 4$, and 5 for $|\mathcal{D}_f| = 5$.

For Office-Home, the domain set is $\mathcal{D} = \{art, clipart, product, real\}$. For $|\mathcal{D}_f| = 1$, we consider $\{art\}$, $\{clipart\}$, $\{product\}$, and $\{real\}$. For $|\mathcal{D}_f| = 2$, we consider $\{art, clipart\}$, $\{art, product\}$, $\{art, real\}$, $\{clipart, product\}$, $\{clipart, real\}$, and $\{product, real\}$. For $|\mathcal{D}_f| = 3$, we consider $\{art, clipart, product\}$, $\{art, clipart, real\}$, $\{art, product, real\}$, and $\{clipart, product, real\}$.

For Mini DomainNet, the domain set is $\mathcal{D} = \{clipart, painting, real, sketch\}$. For $|\mathcal{D}_f| = 1$, we consider $\{clipart\}$, $\{painting\}$, $\{real\}$, and $\{sketch\}$. For $|\mathcal{D}_f| = 2$, we consider $\{clipart, painting\}$, $\{clipart, real\}$, $\{clipart, sketch\}$, $\{painting, real\}$, $\{painting, sketch\}$, and $\{real, sketch\}$. For $|\mathcal{D}_f| = 3$, we consider $\{clipart, painting, real\}$, $\{clipart, painting, sketch\}$, $\{clipart, real, sketch\}$, and $\{painting, real, sketch\}$.

For DomainNet, the domain set is $\mathcal{D} = \{clipart, infograph, painting, quickdraw, real, sketch\}$. For $|\mathcal{D}_f| = 1$, we consider all configurations: $\{clipart\}$, $\{infograph\}$, $\{painting\}$, $\{quickdraw\}$, $\{real\}$, and $\{sketch\}$. For $|\mathcal{D}_f| = 2$, we consider three configurations: $\{clipart, infograph\}$, $\{painting, quickdraw\}$, and $\{real, sketch\}$. For $|\mathcal{D}_f| = 3$, we use $\{clipart, infograph, painting\}$, and for $|\mathcal{D}_f| = 4$, we use $\{clipart, infograph, painting, quickdraw\}$. For $|\mathcal{D}_f| = 5$, we consider five configurations: $\{clipart, infograph, painting, quickdraw, real\}$, $\{clipart, infograph, quickdraw, real, sketch\}$, $\{clipart, infograph, painting, quickdraw, sketch\}$, $\{clipart, painting, quickdraw, real, sketch\}$, and $\{infograph, painting, quickdraw, real, sketch\}$.

D Detailed Retrieval Results

Table 2 in the main paper reports retrieval performance averaged over R@1, R@5, and R@10. Here, we provide the corresponding fine-grained results at each recall level separately. Tables 4 and 5 report *Mem*, *For*, and *H* at R@1, R@5, and R@10 for text-to-image and image-to-text retrieval, respectively, across forget-set sizes $|\mathcal{D}_f| \in \{1, 2, 3, 4, 5\}$. The direction- and recall-specific results are consistent with the aggregate trends reported in the main paper. ADU achieves strong forgetting but substantially degrades retain-domain retrieval, whereas DACE maintains a more balanced retain–forget trade-off. Across recall levels, DACE-H generally achieves the strongest retrieval harmonic mean, while DACE-S favors greater retention and DACE-M provides an intermediate operating point.

Table 4: Text-to-image domain unlearning on DomainNet at R@1, R@5, and R@10, averaged over forget/retain configurations. **Bold**/underlined: best/second-best per metric within $|\mathcal{D}_f|$ groups.

$\|\mathcal{D}_f\|$	Method	Text-to-Image Retrieval								
		Mem↑			*For*↑			*H*↑		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1	Zero-shot CLIP	39.66	64.17	73.21	51.86	26.34	19.39	44.11	33.15	25.97
	ADU (Kawamura et al., 2025)	0.60	1.63	2.35	**99.48**	**98.94**	**98.45**	1.19	3.20	4.59
	CLIPErase (Yang et al., 2025)	15.53	32.18	39.93	95.87	93.39	92.65	26.60	47.70	55.65
	DACE-S (ours)	41.78	67.16	75.74	79.09	65.64	60.84	54.62	65.40	65.46
	DACE-M (ours)	38.98	64.26	73.42	90.86	84.31	81.63	54.38	72.90	77.28
	DACE-H (ours)	**60.26**	**85.47**	**90.80**	97.10	95.80	94.01	**72.94**	**90.30**	**92.36**
2	Zero-shot CLIP	41.15	65.98	74.84	54.77	29.29	21.44	46.93	39.64	31.95
	ADU (Kawamura et al., 2025)	0.73	1.89	2.61	**99.60**	**99.09**	**98.85**	1.44	3.70	5.08
	CLIPErase (Yang et al., 2025)	7.26	15.95	20.16	98.46	97.10	96.47	13.06	26.16	31.89
	DACE-S (ours)	41.48	67.48	76.25	78.40	65.26	60.68	54.19	66.19	67.12
	DACE-M (ours)	34.89	61.03	71.02	90.95	81.75	76.53	50.31	69.82	73.66
	DACE-H (ours)	**58.26**	**82.07**	**87.70**	97.54	94.83	93.33	**72.19**	**87.81**	**90.35**
3	Zero-shot CLIP	34.63	57.48	66.48	48.23	21.35	13.85	40.32	31.13	22.92
	ADU (Kawamura et al., 2025)	0.75	2.20	3.29	99.71	99.54	99.15	1.49	4.30	6.36
	CLIPErase (Yang et al., 2025)	0.01	0.17	0.28	**99.96**	**99.89**	**99.83**	0.02	0.34	0.55
	DACE-S (ours)	35.14	58.76	67.63	78.83	65.54	60.34	48.61	61.97	63.78
	DACE-M (ours)	25.26	50.22	61.86	91.55	82.09	76.55	39.60	62.32	68.42
	DACE-H (ours)	**57.10**	**77.78**	**84.73**	98.94	98.26	97.68	**72.41**	**86.83**	**90.75**
4	Zero-shot CLIP	51.28	81.65	89.91	60.22	37.54	29.65	55.39	51.43	44.60
	ADU (Kawamura et al., 2025)	1.20	3.15	4.60	99.80	99.47	99.22	2.36	6.11	8.79
	CLIPErase (Yang et al., 2025)	0.04	0.13	0.34	**99.98**	**99.95**	**99.90**	0.07	0.25	0.69
	DACE-S (ours)	49.39	81.87	90.67	79.18	63.53	56.55	60.84	71.54	69.66
	DACE-M (ours)	38.88	71.36	82.57	92.46	84.43	79.85	54.74	77.35	81.19
	DACE-H (ours)	**79.42**	**96.09**	**98.26**	99.13	97.46	96.74	**88.19**	**96.77**	**97.49**
5	Zero-shot CLIP	56.21	84.14	90.56	61.58	37.70	28.78	58.54	52.03	43.66
	ADU (Kawamura et al., 2025)	1.38	3.61	5.05	99.87	99.65	99.52	2.71	6.91	9.51
	CLIPErase (Yang et al., 2025)	0.02	0.15	0.34	**99.99**	**99.96**	**99.94**	0.04	0.31	0.67
	DACE-S (ours)	48.25	79.67	87.68	89.23	78.61	73.46	62.29	78.74	79.52
	DACE-M (ours)	42.51	73.91	83.85	95.64	89.07	84.55	58.66	80.63	84.07
	DACE-H (ours)	**79.23**	**93.79**	**96.06**	98.79	97.48	96.25	**87.80**	**95.50**	**96.08**

Table 5: Image-to-text domain unlearning on DomainNet at R@1, R@5, and R@10, averaged over forget/retain configurations. **Bold**/underlined: best/second-best per metric within $|\mathcal{D}_f|$ groups.

$\|\mathcal{D}_f\|$	Method	Image-to-Text Retrieval								
		Mem↑			*For*↑			*H*↑		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1	Zero-shot CLIP	40.98	66.21	74.12	52.27	26.72	20.44	45.01	33.34	26.74
	ADU (Kawamura et al., 2025)	1.09	2.22	2.82	**99.67**	**99.14**	**98.69**	2.16	4.34	5.47
	CLIPErase (Yang et al., 2025)	16.09	34.02	42.37	96.15	93.40	92.42	27.31	49.57	57.84
	DACE-S (ours)	41.35	67.57	75.73	77.11	61.85	56.60	53.73	62.79	61.66
	DACE-M (ours)	40.64	66.19	74.67	93.43	88.42	86.24	56.48	75.66	80.03
	DACE-H (ours)	**61.72**	**84.61**	**89.38**	96.67	93.09	90.58	**74.91**	**88.62**	**89.91**
2	Zero-shot CLIP	42.39	67.82	75.39	53.93	28.48	21.66	47.38	38.88	31.98
	ADU (Kawamura et al., 2025)	1.34	3.09	3.79	**99.67**	**99.23**	**98.86**	2.63	5.96	7.25
	CLIPErase (Yang et al., 2025)	8.38	17.49	22.21	98.58	97.45	96.99	14.75	27.82	33.84
	DACE-S (ours)	40.86	67.67	76.01	78.54	63.87	58.60	53.69	65.15	65.12
	DACE-M (ours)	37.48	63.61	72.82	93.12	86.11	82.79	53.33	73.08	77.44
	DACE-H (ours)	**58.09**	**78.49**	**84.68**	96.71	94.39	92.27	**72.48**	**85.63**	**88.26**
3	Zero-shot CLIP	38.24	60.33	67.19	49.78	21.36	14.17	43.25	31.55	23.40
	ADU (Kawamura et al., 2025)	1.67	3.26	3.84	99.59	99.20	99.06	3.28	6.31	7.40
	CLIPErase (Yang et al., 2025)	0.06	0.27	0.42	**99.99**	**99.89**	**99.79**	0.12	0.53	0.84
	DACE-S (ours)	34.88	58.90	67.36	78.57	62.57	56.00	48.31	60.68	61.16
	DACE-M (ours)	29.23	54.69	64.25	93.42	86.28	82.03	44.53	66.94	72.06
	DACE-H (ours)	**53.62**	**74.30**	**81.26**	99.71	99.03	98.45	**69.74**	**84.90**	**89.03**
4	Zero-shot CLIP	56.38	85.91	92.36	61.39	37.78	30.43	58.78	52.48	45.77
	ADU (Kawamura et al., 2025)	2.54	4.64	6.01	99.84	99.62	99.42	4.95	8.86	11.34
	CLIPErase (Yang et al., 2025)	0.04	0.24	0.34	**99.98**	**99.95**	**99.84**	0.07	0.47	0.69
	DACE-S (ours)	49.76	81.52	88.84	84.18	71.53	65.90	62.54	76.20	75.67
	DACE-M (ours)	44.26	77.09	85.41	94.39	87.94	84.94	60.26	82.16	85.17
	DACE-H (ours)	**74.20**	**92.32**	**95.65**	99.64	98.77	98.41	**85.06**	**95.43**	**97.01**
5	Zero-shot CLIP	55.87	84.21	90.24	60.33	35.81	27.98	57.72	50.21	42.70
	ADU (Kawamura et al., 2025)	4.26	7.48	8.83	99.80	99.56	99.38	8.13	13.76	16.03
	CLIPErase (Yang et al., 2025)	0.02	0.19	0.41	**100.00**	**99.97**	**99.94**	0.04	0.38	0.82
	DACE-S (ours)	46.97	78.02	86.00	94.06	86.62	82.70	62.44	81.91	84.15
	DACE-M (ours)	45.93	76.85	84.57	97.82	94.64	92.54	62.28	84.60	88.19
	DACE-H (ours)	**68.22**	**90.14**	**93.74**	99.62	98.97	98.30	**80.20**	**94.23**	**95.89**

E Additional DDL Ablation Results

To complement the harmonic-mean (H) results reported in Figure 2 of the main paper, we report the corresponding *Mem* and *For* components in Figures 4 and 5, respectively. These results provide a more detailed view of how the different DDL variants affect the retain–forget trade-off across classification, text-to-image retrieval, and image-to-text retrieval.



Figure 4: Ablation of DDL variants on *Mem* metric on DomainNet with $|\mathcal{D}_f| = 1$, averaged over forget-domain choices. DACE-S, DACE-M, and DACE-H are evaluated without DDL, with G DDL, and with RF DDL across classification and bidirectional retrieval.

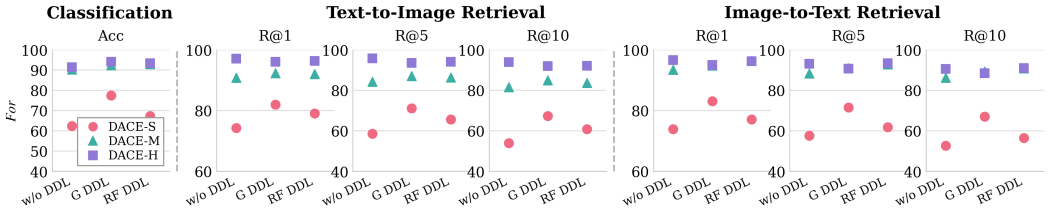


Figure 5: Ablation of DDL variants on *For* metric on DomainNet with $|\mathcal{D}_f| = 1$, averaged over forget-domain choices. DACE-S, DACE-M, and DACE-H are evaluated without DDL, with G DDL, and with RF DDL across classification and bidirectional retrieval.

F Ablation of Training Objectives

Tables 6, 7, and 8 provide a detailed ablation of the training objectives on DomainNet with one forget domain ($|\mathcal{D}_f| = 1$). We separately examine the roles of the paired contrastive retain loss $\mathcal{L}_r$, the alignment-editing forgetting loss $\mathcal{L}_f^{\psi}$, and the domain-disentanglement loss $\mathcal{L}_{\mathrm{DDL}}$. Here, $\mathcal{L}_f^{\psi}$ corresponds to similarity suppression (DACE-S), mismatch reassignment (DACE-M), or hard-negative margin forgetting (DACE-H). We additionally consider two auxiliary forms of retain supervision: he cross-entropy (CE) retain loss, which applies cross-entropy supervision to retain-domain samples, and $\mathcal{L}_r^{\mathrm{CP}}$, the class-prompt (CP) retain loss, which preserves classification alignment with class-prompt embeddings. These objectives supplement, rather than replace, the paired contrastive retain loss. In the tables, "CE retain loss" and "CP retain loss" denote these two additional terms, respectively.

Retain and forgetting objectives. The retain-only and forget-only configurations illustrate the need to balance preservation and forgetting. Using $\mathcal{L}_f^{\psi}$ alone produces near-complete forgetting but severely degrades retain-domain utility, whereas $\mathcal{L}_r$ alone largely preserves retain-domain alignment but provides insufficient forgetting. Combining $\mathcal{L}_r$ and $\mathcal{L}_f^{\psi}$ yields a substantially better retain–forget trade-off even in the absence of $\mathcal{L}_{\mathrm{DDL}}$. In classification, DACE-M without DDL achieves the highest harmonic mean of 68.06. For retrieval, DACE-H without DDL achieves $H@1/5/10$ values of $72.94/90.30/92.36$ for text-to-image retrieval and $74.91/88.62/89.91$ for image-to-text retrieval. These results show that the proposed alignment-editing objectives can induce effective forgetting without relying on domain disentanglement.

Role of domain disentanglement. Adding $\mathcal{L}_{\mathrm{DDL}}$ primarily shifts the operating point toward stronger forgetting. In classification, the forgetting score increases from 62.49 to 67.37 for DACE-S,

from 90.29 to 92.96 for DACE-M, and from 91.54 to 93.38 for DACE-H. The corresponding effect on the harmonic mean, however, depends on the forgetting operator. DACE-S benefits more from explicit domain disentanglement because similarity suppression provides a comparatively weaker forgetting signal. DACE-M and DACE-H already achieve strong forgetting without DDL; for these variants, DDL further increases forgetting but can reduce retain performance, resulting in a smaller or negative change in the overall harmonic mean. Thus, $\mathcal{L}_{\mathrm{DDL}}$ controls the retain–forget operating point rather than serving as the primary source of the forgetting signal.

Additional retain supervision. Finally, we evaluate whether augmenting the paired contrastive retain objective with $\mathcal{L}_r^{\mathrm{CE}}$ or $\mathcal{L}_r^{\mathrm{CP}}$ further improves the retain–forget trade-off. The corresponding classification and direction-specific retrieval results are reported below, allowing us to assess whether additional task-oriented retain supervision provides benefits beyond the original cross-modal retain objective.

Table 6: Loss-component ablation on DomainNet ($|\mathcal{D}_f| = 1$), with performance aggregated over all available configurations and reported as the mean. Within each metric column, **bold** indicates the best result and underlined indicates the second-best result.

Method	$\mathcal{L}_r$	$\mathcal{L}_f$	$\mathcal{L}_{\mathrm{DDL}}$	$\mathcal{L}_r^{\mathrm{CE}}$	$\mathcal{L}_{r,\mathrm{cls}}$	Classification		
						Mem↑	*For*↑	*H*↑
Retain-only DACE	✓	–	–	–	–	55.68	46.20	48.92
Forget-only DACE-S	–	✓	–	–	–	0.36	**99.76**	0.72
Forget-only DACE-M	–	✓	–	–	–	0.35	99.73	0.69
Forget-only DACE-H	–	✓	–	–	–	1.17	99.71	2.24
DACE-S w/o DDL	✓	✓	–	–	–	54.97	62.49	56.88
DACE-M w/o DDL	✓	✓	–	–	–	54.67	90.29	68.06
DACE-H w/o DDL	✓	✓	–	–	–	51.57	91.54	65.86
DACE-S	✓	✓	✓	–	–	56.11	67.37	60.05
DACE-M	✓	✓	✓	–	–	51.02	92.96	65.75
DACE-H	✓	✓	✓	–	–	50.03	93.38	65.01
DACE-S + CE retain	✓	✓	✓	✓	–	58.78	28.89	37.78
DACE-M + CE retain	✓	✓	✓	✓	–	60.14	85.97	**70.77**
DACE-H + CE retain	✓	✓	✓	✓	–	59.03	84.68	69.55
DACE-S + CP retain	✓	✓	✓	–	✓	**62.15**	60.00	58.56
DACE-M + CP retain	✓	✓	✓	–	✓	57.67	85.10	68.75
DACE-H + CP retain	✓	✓	✓	–	✓	58.82	86.80	70.11

Table 7: Loss-component ablation for Text-to-Image retrieval on DomainNet ($|\mathcal{D}_f| = 1$), with performance aggregated over all available configurations and reported as the mean. Within each metric column, **bold** indicates the best result and underlined indicates the second-best result.

Method	$\mathcal{L}_r$	$\mathcal{L}_f$	$\mathcal{L}_{\mathrm{DDL}}$	$\mathcal{L}_r^{\mathrm{CE}}$	$\mathcal{L}_{r,\mathrm{cls}}$	Text-to-Image Retrieval								
						Mem↑			*For*↑			*H*↑		
						R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Retain-only DACE	✓	–	–	–	–	**72.26**	**88.78**	**92.74**	23.00	11.69	8.41	30.25	16.41	12.56
Forget-only DACE-S	–	✓	–	–	–	0.13	0.34	0.44	**100.00**	**99.96**	**99.93**	0.26	0.66	0.87
Forget-only DACE-M	–	✓	–	–	–	0.10	0.25	0.38	99.97	99.91	99.82	0.19	0.49	0.75
Forget-only DACE-H	–	✓	–	–	–	1.64	3.12	3.90	**100.00**	99.57	99.28	3.01	5.36	6.55
DACE-S w/o DDL	✓	✓	–	–	–	41.82	68.19	77.06	74.31	58.67	54.05	53.32	60.18	58.81
DACE-M w/o DDL	✓	✓	–	–	–	38.98	64.26	73.42	90.86	84.31	81.63	54.38	72.90	77.28
DACE-H w/o DDL	✓	✓	–	–	–	60.26	85.47	90.80	97.10	95.80	94.01	**72.94**	**90.30**	**92.36**
DACE-S	✓	✓	✓	–	–	41.78	67.16	75.74	79.09	65.64	60.84	54.62	65.40	65.46
DACE-M	✓	✓	✓	–	–	35.30	60.11	69.42	92.09	86.39	83.71	50.78	70.80	75.88
DACE-H	✓	✓	✓	–	–	55.13	81.96	87.54	96.38	94.10	92.22	68.99	87.55	89.79
DACE-S + CE retain	✓	✓	✓	✓	–	39.10	63.75	72.48	63.81	37.51	28.53	48.42	45.96	38.57
DACE-M + CE retain	✓	✓	✓	✓	–	35.75	60.50	69.69	89.08	81.66	78.29	51.02	69.50	73.73
DACE-H + CE retain	✓	✓	✓	✓	–	58.41	83.14	88.52	95.02	92.17	89.80	71.50	87.37	89.07
DACE-S + CP retain	✓	✓	✓	–	✓	41.14	67.19	76.07	75.80	61.60	57.41	53.00	61.33	60.95
DACE-M + CP retain	✓	✓	✓	–	✓	34.23	58.33	67.64	88.75	80.03	76.59	49.31	67.42	71.80
DACE-H + CP retain	✓	✓	✓	–	✓	58.99	82.88	88.69	94.93	92.32	90.24	72.30	87.31	89.40

Table 8: Loss-component ablation for Image-to-Text retrieval on DomainNet ($|\mathcal{D}_f| = 1$), with performance aggregated over all available configurations and reported as the mean. Within each metric column, **bold** indicates the best result and underlined indicates the second-best result.

Method	$\mathcal{L}_r$	$\mathcal{L}_f$	$\mathcal{L}_{\text{DDL}}$	$\mathcal{L}_r^{\text{CE}}$	$\mathcal{L}_{r,\text{cls}}$	Image-to-Text Retrieval								
						Mem↑			*For*↑			*H*↑		
						R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Retain-only DACE	✓	–	–	–	–	**72.63**	**87.89**	**91.78**	25.71	13.67	10.29	32.90	19.06	15.27
Forget-only DACE-S	–	✓	–	–	–	0.05	0.12	0.21	**99.97**	**99.82**	**99.68**	0.10	0.25	0.41
Forget-only DACE-M	–	✓	–	–	–	0.01	0.06	0.09	99.96	**99.82**	99.67	0.02	0.11	0.19
Forget-only DACE-H	–	✓	–	–	–	0.88	1.99	2.69	99.85	99.37	98.55	1.68	3.63	4.83
DACE-S w/o DDL	✓	✓	–	–	–	41.15	68.22	76.99	73.98	57.65	52.67	52.69	59.70	57.94
DACE-M w/o DDL	✓	✓	–	–	–	40.64	66.19	74.67	93.43	88.42	86.24	56.48	75.66	80.03
DACE-H w/o DDL	✓	✓	–	–	–	61.72	84.61	89.38	96.67	93.09	90.58	**74.91**	**88.62**	**89.91**
DACE-S	✓	✓	✓	–	–	41.35	67.57	75.73	77.11	61.85	56.60	53.73	62.79	61.66
DACE-M	✓	✓	✓	–	–	32.70	58.33	67.47	96.44	92.82	90.86	48.58	71.45	77.33
DACE-H	✓	✓	✓	–	–	58.40	80.40	86.03	96.33	93.43	91.01	72.27	86.35	88.39
DACE-S + CE retain	✓	✓	✓	✓	–	39.13	64.71	73.21	57.72	28.92	21.20	46.22	37.22	29.67
DACE-M + CE retain	✓	✓	✓	✓	–	34.05	60.26	69.91	95.03	88.93	86.64	50.13	71.84	77.38
DACE-H + CE retain	✓	✓	✓	✓	–	59.82	81.88	87.11	95.17	90.53	87.19	73.16	85.94	87.07
DACE-S + CP retain	✓	✓	✓	–	✓	40.41	67.19	75.65	75.12	58.40	53.18	52.20	59.27	57.87
DACE-M + CP retain	✓	✓	✓	–	✓	30.97	55.82	65.67	94.26	87.73	84.69	46.47	68.14	73.92
DACE-H + CP retain	✓	✓	✓	–	✓	57.13	81.50	86.95	95.65	91.49	88.30	70.91	86.16	87.55

G Additional Ablation Results for Training Sample Sensitivity

To provide a more detailed understanding of the retain–forgetting trade-off, we report the individual components of the harmonic mean, namely *Mem* and *For*, as a function of the number of shots. The corresponding results are shown in Figures 6 and 7. From Figure 6, we observe that forgetting performance improves as the number of shots increases across all tasks. DACE-H consistently achieves the strongest forgetting, maintaining high For scores even in low-shot settings, while further benefiting from additional data. DACE-M also demonstrates strong and stable forgetting, whereas DACE-S exhibits lower For values, reflecting its conservative strategy of suppressing, rather than aggressively modifying, image–text alignment. Figure 7 shows the corresponding memorization performance. As expected, increasing the number of shots leads to improved retain-domain performance for all methods. DACE-S consistently achieves higher memorization, highlighting its preservation-oriented behavior. In contrast, DACE-M and DACE-H trade some memorization for stronger forgetting, particularly in low-shot regimes. However, this gap decreases as more data becomes available, indicating that additional supervision helps mitigate the retention cost of stronger unlearning. Overall, these results provide a more complete view of the trade-off captured by *H* in the main paper. They confirm that the differences between DACE variants stem from distinct alignment-editing behaviors: DACE-S prioritizes retention, DACE-M balances retention and forgetting, and DACE-H emphasizes strong forgetting. This analysis further supports the robustness of the proposed framework across both tasks and data regimes.



Figure 6: *For* versus the number of shots on DomainNet with $|\mathcal{D}_f| = 1$, averaged over all possible choices of the forget domain, for classification and bidirectional retrieval.

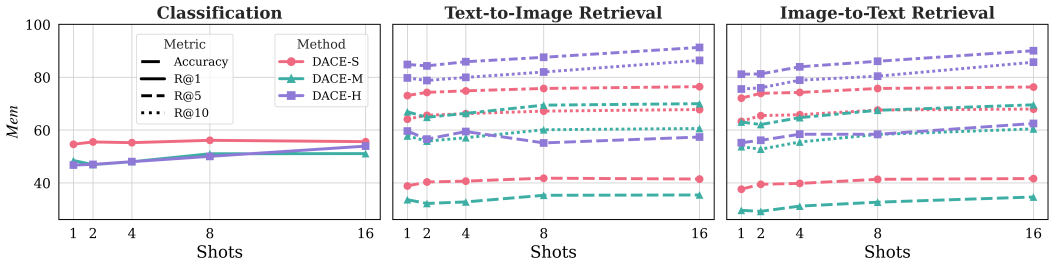
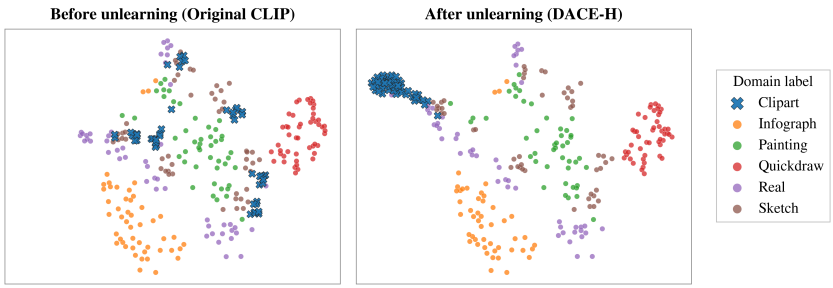


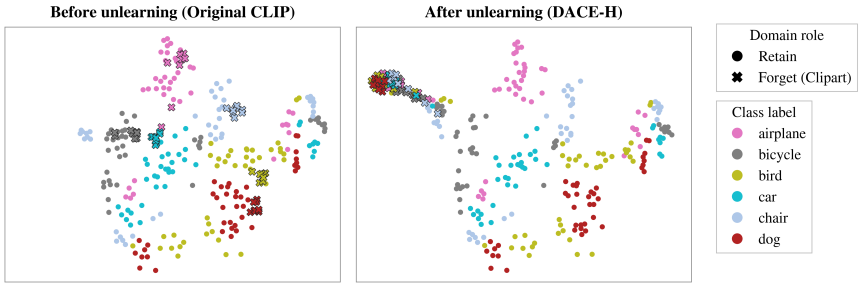
Figure 7: *Mem* versus the number of shots on DomainNet with $|\mathcal{D}_f| = 1$, averaged over all possible choices of the forget domain, for classification and bidirectional retrieval.

H Representation-Level Analysis

We visualize the representation-level effects of DACE-H before and after unlearning, using Clipart as the forget domain. For clarity and balanced visualization, we construct a subset spanning six semantic classes (Airplane, Bicycle, Bird, Car, Chair, and Dog) across all six DomainNet domains. We sample up to eight images per class from each domain, yielding 286 images in total. Specifically, each class–domain pair contains eight images except for the Car class in Painting, for which six images are available in the evaluation subset. This results in 48 images from each domain except Painting, which contains 46 images. The same image subset is used for all before–after comparisons. Figures 8a, 9a, and 10a colors image embeddings by domain. After unlearning, Clipart representations undergo a substantial reorganization and become more concentrated relative to the retain-domain representations. Figures 8b, 9b, and 10b shows the same image embeddings colored by semantic class. Before unlearning, Clipart samples exhibit visually distinguishable class-dependent structure, whereas after unlearning, samples from different classes appear more intermixed within the reorganized Clipart representation.

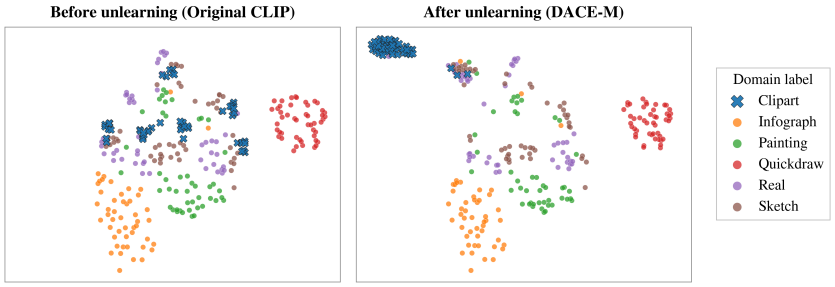


(a) t-SNE visualization of image embeddings before and after DACE-H unlearning of the Clipart domain. Colors indicate domains, and crosses denote Clipart images.

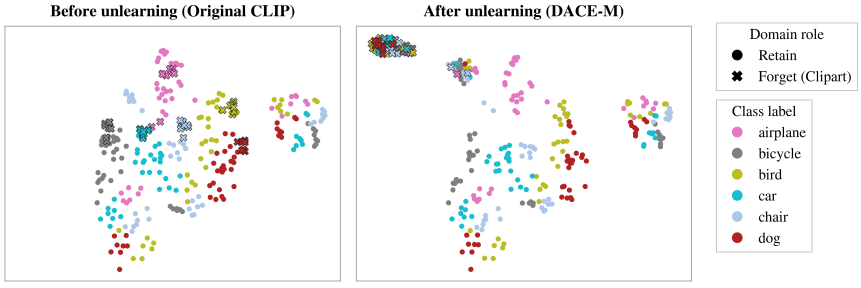


(b) t-SNE visualization of semantic class structure before and after DACE-H unlearning of the Clipart domain. Colors indicate classes; circles and crosses denote retain- and forget-domain images, respectively.

Figure 8: Representation-level analysis before and after DACE-H unlearning of the Clipart domain.

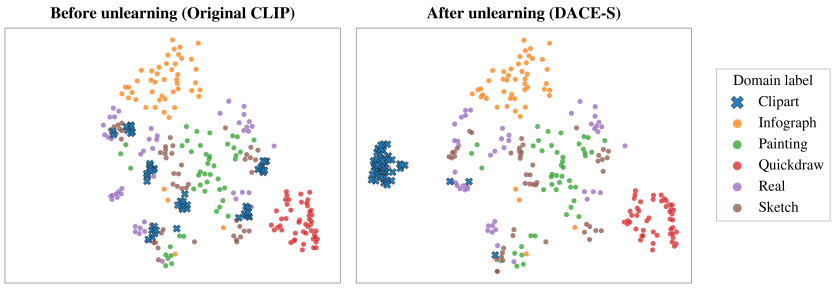


(a) t-SNE visualization of image embeddings before and after DACE-M unlearning of the Clipart domain. Colors indicate domains, and crosses denote Clipart images.

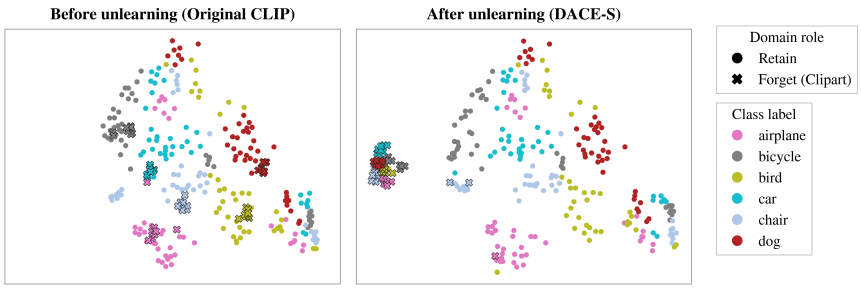


(b) t-SNE visualization of semantic class structure before and after DACE-M unlearning of the Clipart domain. Colors indicate classes; circles and crosses denote retain- and forget-domain images, respectively.

Figure 9: Representation-level analysis before and after DACE-M unlearning of the Clipart domain.



(a) t-SNE visualization of image embeddings before and after DACE-S unlearning of the Clipart domain. Colors indicate domains, and crosses denote Clipart images.



(b) t-SNE visualization of semantic class structure before and after DACE-S unlearning of the Clipart domain. Colors indicate classes; circles and crosses denote retain- and forget-domain images, respectively.

Figure 10: Representation-level analysis before and after DACE-S unlearning of the Clipart domain.

I DISENTANGLING VISUAL STRUCTURE FROM CROSS-MODAL ALIGNMENT

A decrease in forget-domain classification or retrieval performance does not by itself characterize how unlearning modifies the underlying multimodal representation. In particular, changes in the visual organization of classes need not coincide with changes in image–text alignment. We therefore analyze these two properties separately: (i) class separation within the visual embedding space and (ii) alignment between images and their corresponding class–text representations. This analysis complements the task-level Mem/For evaluation by providing a representation-level view of how different unlearning objectives modify the shared embedding space. We compare Original CLIP, ADU, and DACE-S/M/H using Clipart as the forget domain. The forget-domain analysis uses the same 10,663 held-out Clipart images spanning 329 classes for all methods. We additionally compute the visual-geometry diagnostic separately on each of the five retained DomainNet domains using identical evaluation images across methods.

Visual semantic structure. We first characterize the visual geometry within the forgotten domain. For the Clipart forget set, let z_i denote the normalized image embedding of sample i and y_i its class label. All quantities below are computed exclusively over the 10,663 held-out Clipart images spanning 329 classes. We define the sets of same-class and different-class image pairs as

$$\mathcal{P}_{\text{same}} = \{(i,j) : i \neq j,\ y_i = y_j\}, \qquad \mathcal{P}_{\text{diff}} = \{(i,j) : y_i \neq y_j\}. \tag{44}$$

We then measure

$$S_{\text{same}} = \frac{1}{|\mathcal{P}_{\text{same}}|} \sum_{(i,j)\in\mathcal{P}_{\text{same}}} z_i^\top z_j, \quad S_{\text{diff}} = \frac{1}{|\mathcal{P}_{\text{diff}}|} \sum_{(i,j)\in\mathcal{P}_{\text{diff}}} z_i^\top z_j, \quad G_{\text{sem}} = S_{\text{same}} - S_{\text{diff}}. \tag{45}$$

Here, G_{sem} measures average class separation in the visual embedding space: larger values indicate a larger difference between within-class and between-class similarity. Self-pairs are excluded and all image pairs receive equal weight. We compute these statistics exactly using class-wise feature sums without explicitly constructing the full pairwise similarity matrix.

Cross-modal alignment. To separately characterize image–text correspondence, let $w_c = \frac{f_T(p_c)}{\|f_T(p_c)\|_2}$ denote the normalized text embedding of class c, where p_c is the class prompt used by the corresponding model's classification head, including learned context when applicable. For N evaluation images, we measure correct-class image–text alignment as

$$S_{\text{IT}}^{\text{native}} = \frac{1}{N} \sum_{i=1}^{N} z_i^\top w_{y_i}. \tag{46}$$

Unlike G_{sem}, which depends only on image embeddings, $S_{\text{IT}}^{\text{native}}$ measures correct-class cross-modal alignment under the model's native text representation. The two quantities therefore characterize complementary visual and cross-modal properties of the learned representation. For ease of comparison, we additionally report the relative change in the visual semantic gap with respect to Original CLIP: $\text{GapReduction}^{(m)} = 100\left(1 - \frac{G_{\text{sem}}^{(m)}}{G_{\text{sem}}^{(\text{CLIP})}}\right)$. This quantity measures the relative reduction of this particular geometric statistic and should not be interpreted as the fraction of semantic information removed from the representation.

Forget-domain representation. Table 9 shows that visual class geometry and correct-class image–text alignment respond differently to the unlearning objectives. DACE-S retains 59.0% of Original CLIP's visual semantic gap (0.0746 versus 0.1263), while producing the largest reduction in correct-class image–text alignment, from 0.2719 to 0.1661 (38.9%). Thus, for DACE-S, correct-class image–text alignment decreases substantially while a comparatively larger fraction of the original visual class separation remains. The other DACE operators exhibit distinct representation-level behaviors. DACE-H reduces the semantic gap by 71.2% and correct-class alignment by 28.4%, whereas DACE-M exhibits a substantially larger reduction in the visual semantic gap (90.7%). ADU similarly produces a large reduction in the visual gap (89.5%). Importantly, the ordering of methods according to G_{sem} differs from their ordering according to $S_{\text{IT}}^{\text{native}}$, indicating that changes in visual class geometry and changes in correct-class cross-modal alignment are not equivalent effects.

Table 9: Visual semantic structure and correct-class image–text alignment on the forgotten Clipart domain. All methods are evaluated on the same 10,663 images spanning 329 classes. Similarities are raw cosine values and results are point estimates for one seed.

Method	S_{same}	S_{diff}	G_{sem}	Gap reduction (%)	$S_{\text{IT}}^{\text{native}}$
Original CLIP	0.7342	0.6079	0.1263	—	0.2719
ADU	0.8296	0.8164	0.0132	89.5	0.2209
DACE-S	0.7178	0.6432	0.0746	41.0	0.1661
DACE-M	0.8920	0.8803	0.0118	90.7	0.2025
DACE-H	0.7646	0.7282	0.0364	71.2	0.1947

Table 10: Visual semantic gap G_{sem} on retained DomainNet domains. Each column uses identical evaluation images across methods.

Domain	Original CLIP	ADU	DACE-S	DACE-M	DACE-H
Infograph	0.1000	0.0472	0.0836	0.0533	0.0680
Painting	0.1417	0.0972	0.1481	0.0868	0.1086
Quickdraw	0.0417	0.0167	0.0474	0.0361	0.0395
Real	0.2070	0.0724	0.2399	0.1033	0.1560
Sketch	0.0980	0.0210	0.1139	0.0481	0.0674

Retain-domain representation. Table 10 extends the visual-geometry analysis to the five retained domains. DACE-S retains the largest semantic gap among the evaluated unlearning methods, exceeding Original CLIP in four of the five retained domains. DACE-H generally occupies an intermediate regime, whereas ADU and DACE-M induce larger changes in visual geometry. Together with the forget-domain results, these patterns indicate that the three DACE objectives induce distinct changes in representation geometry rather than simply different magnitudes of the same geometric effect.

Overall, these results distinguish complementary aspects of domain unlearning that are not captured by task-level performance alone: visual class geometry and cross-modal alignment. Most notably, DACE-S reduces correct-class image–text alignment by 38.9% on the forgotten Clipart domain while retaining 59.0% of Original CLIP's visual semantic gap. DACE-M, in contrast, exhibits a substantially larger reduction in the visual semantic gap. These results show that changes in visual class geometry and correct-class cross-modal alignment need not occur to the same extent under different unlearning objectives. The representation-level diagnostics therefore complement Mem/For by characterizing how different DACE operators modify the shared visual–language representation.

J GENERALIZATION TO SIGLIP

To evaluate whether DACE generalizes beyond CLIP, we extend the framework to SigLIP and evaluate it on DomainNet using the same single-domain unlearning protocol and paired image–caption supervision as in the main experiments. The overall DACE formulation remains unchanged:

$$\mathcal{L}_{\text{DACE}} = \lambda_r \mathcal{L}_r(\mathcal{B}_r) + \lambda_f \mathcal{L}_f^{\psi}(\mathcal{B}_f) + \mathcal{L}_{\text{DDL}}(\mathcal{B}), \tag{47}$$

where $\mathcal{B}_r$ and $\mathcal{B}_f$ denote retain domain samples and forget domain samples, respectively. The difference is that CLIP uses a softmax-based contrastive objective, whereas SigLIP independently predicts whether each image–text pair is matched. We therefore replace the CLIP-specific contrastive terms with SigLIP-native counterparts that preserve the intended behavior of each DACE operator.

SigLIP adaptation. Let $\mathcal{E}_V$ and $\mathcal{E}_T$ denote the pretrained SigLIP vision and text encoders. Given an image x_i and text t_j, their normalized embeddings are $\mathbf{z}_i^V = \frac{\mathcal{E}_V(x_i)}{\|\mathcal{E}_V(x_i)\|_2}$, $\mathbf{z}_j^T = \frac{\mathcal{E}_T(t_j)}{\|\mathcal{E}_T(t_j)\|_2}$. Both pretrained encoders and SigLIP's global logit-scale and logit-bias parameters are frozen. Since the Hugging Face SigLIP implementation does not expose the same deep visual-prompt insertion points as our modified CLIP backbone, we employ a lightweight residual bottleneck adapter after the frozen

visual projection:

$$\widetilde{\mathbf{z}}_i^V = \mathrm{Norm}\left(\mathbf{z}_i^V + \gamma \mathbf{W}_{\mathrm{up}} \sigma\left(\mathbf{W}_{\mathrm{down}} \mathbf{z}_i^V\right)\right), \tag{48}$$

where σ is the GELU activation, γ is a learnable scalar, and $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. Only this adapter and the domain classifier are optimized. The SigLIP logit for image i and text j is

$$a_{ij} = \exp(s)\widetilde{\mathbf{z}}_i^{V\top} \mathbf{z}_j^T + b, \tag{49}$$

where s and b are the frozen pretrained logit-scale and logit-bias parameters. Freezing them prevents the model from satisfying the forgetting objective through global logit recalibration.

SigLIP retain loss. For paired image–caption supervision, let $\mathcal{P}_r = \{(i,i) : i \in \mathcal{B}_r\}$ denote the set of positive pairs and let $\mathcal{N}_r = \{(i,j) : i,j \in \mathcal{B}_r,\ i \neq j\}$ denote the set of negative pairs. We assign equal importance to the positive and negative sets:

$$\mathcal{L}_r^{\mathrm{SL}}(\mathcal{B}_r) = \frac{1}{2}\left[\frac{1}{|\mathcal{P}_r|}\sum_{(i,j)\in\mathcal{P}_r} \mathrm{softplus}(-a_{ij}) + \frac{1}{|\mathcal{N}_r|}\sum_{(i,j)\in\mathcal{N}_r} \mathrm{softplus}(a_{ij})\right]. \tag{50}$$

For one-caption-per-image supervision, this becomes

$$\mathcal{L}_r^{\mathrm{SL}}(\mathcal{B}_r) = \frac{1}{2}\left[\frac{1}{|\mathcal{B}_r|}\sum_{i\in\mathcal{B}_r} \mathrm{softplus}(-a_{ii}) + \frac{1}{|\mathcal{B}_r|(|\mathcal{B}_r|-1)}\sum_{\substack{i,j\in\mathcal{B}_r \\ i\neq j}} \mathrm{softplus}(a_{ij})\right]. \tag{51}$$

Separately averaging the positive and negative sets prevents the $|\mathcal{B}_r|(|\mathcal{B}_r|-1)$ negative pairs from overwhelming the $|\mathcal{B}_r|$ positive pairs. When class prompts are used, all same-class image–text pairs are included in $\mathcal{P}_r$, and the remaining pairs form $\mathcal{N}_r$.

SigLIP DACE-S. DACE-S suppresses the original image–text association on the forget domain. Its SigLIP-native forgetting objective assigns a negative target to every correct forget pair:

$$\mathcal{L}_f^{\mathrm{S,SL}}(\mathcal{B}_f) = \frac{1}{|\mathcal{B}_f|}\sum_{i\in\mathcal{B}_f} \mathrm{softplus}(a_{ii}). \tag{52}$$

Minimizing equation 52 pushes the correct-pair logit toward a negative value and consequently reduces its sigmoid matching probability. This is the pairwise-sigmoid counterpart of the cosine-similarity suppression in Eq. (6) of the main paper.

SigLIP DACE-M. DACE-M suppresses the original association and redirects each forgotten image toward a mismatched text. For every $i \in \mathcal{B}_f$, we deterministically select a cyclic mismatch $\pi(i) \neq i$. Under class-prompt supervision, the mismatch is additionally required to belong to a different class.

The image-to-text forgetting objective is

$$\mathcal{L}_{f,V\to T}^{\mathrm{M,SL}} = \frac{1}{2|\mathcal{B}_f|}\sum_{i\in\mathcal{B}_f}\left[\mathrm{softplus}(a_{ii}) + \mathrm{softplus}\left(-a_{i,\pi(i)}\right)\right]. \tag{53}$$

The first term assigns a negative target to the original pair, while the second assigns a positive target to the selected mismatch. The factor $1/2$ averages these two components, keeping the scale of DACE-M comparable to the single-term DACE-S objective. For paired image–caption supervision, we analogously define the text-to-image term

$$\mathcal{L}_{f,T\to V}^{\mathrm{M,SL}} = \frac{1}{2|\mathcal{B}_f|}\sum_{i\in\mathcal{B}_f}\left[\mathrm{softplus}(a_{ii}) + \mathrm{softplus}\left(-a_{\pi(i),i}\right)\right]. \tag{54}$$

The final symmetric DACE-M objective is

$$\mathcal{L}_f^{\mathrm{M,SL}} = \frac{1}{2}\left(\mathcal{L}_{f,V\to T}^{\mathrm{M,SL}} + \mathcal{L}_{f,T\to V}^{\mathrm{M,SL}}\right). \tag{55}$$

For class-prompt supervision, only the image-to-text direction is used.

SigLIP DACE-H. DACE-H is retained in the normalized embedding space so that its margin cannot be satisfied through changes to the global SigLIP scale or bias. Define

$$q_{ij} = \mathrm{sim}\left(\widetilde{\mathbf{z}}_i^V, \mathbf{z}_j^T\right) = \widetilde{\mathbf{z}}_i^{V\top}\mathbf{z}_j^T. \tag{56}$$

For each forget-domain image, the hardest incorrect text is

$$h_t(i) = \arg\max_{\substack{j \in \mathcal{B}_f \\ j \neq i}} q_{ij}. \tag{57}$$

The image-to-text DACE-H objective is

$$\mathcal{L}^{\mathrm{H,SL}}_{f,V\to T} = \frac{1}{|\mathcal{B}_f|} \sum_{i \in \mathcal{B}_f} \max\left(0,\, q_{ii} - q_{i,h_t(i)} + m\right), \tag{58}$$

where $m > 0$ is the margin. For paired captions, we also compute the analogous text-to-image term and use their average:

$$\mathcal{L}^{\mathrm{H,SL}}_{f} = \frac{1}{2}\left(\mathcal{L}^{\mathrm{H,SL}}_{f,V\to T} + \mathcal{L}^{\mathrm{H,SL}}_{f,T\to V}\right). \tag{59}$$

For class-prompt supervision, all same-class targets are excluded from the hard-negative set, and only the image-to-text direction is used. This formulation is consistent with Eqs. (10)–(14) of the main paper.

Domain disentanglement. We apply the retain–forget-aware DDL variant from the main paper to the adapted SigLIP visual embeddings. Let $r_i \in \{0, 1\}$ indicate whether sample i belongs to a retained or forget domain. The DDL objective remains

$$\mathcal{L}^{\mathrm{RF}}_{\mathrm{DDL}}(\mathcal{B}) = \alpha \mathcal{L}^{\mathrm{RF}}_{\mathrm{CE}}(\mathcal{B}) - \beta\, \mathrm{MMD}^2_{\mathrm{RF}}(\mathcal{B}), \tag{60}$$

where the domain classification and MMD terms are computed using the binary retain–forget partition. We preserve the signs and weighting of the original DDL objective.

Table 11: SigLIP classification on DomainNet, averaged over the same available forget-domain configurations for all methods at each $|\mathcal{D}_f|$.

$\lvert\mathcal{D}_f\rvert$	Method	Classification		
		Mem Acc.↑	*For* Acc.↑	*H* Acc.↑
1	Zero-shot SigLIP	**68.10**	31.90	39.96
	DACE-S	54.19	81.81	64.96
	DACE-M	52.82	**87.15**	65.41
	DACE-H	57.46	79.11	**66.34**
2	Zero-shot SigLIP	**68.10**	31.90	42.14
	DACE-S	50.79	76.12	60.74
	DACE-M	52.38	**79.96**	62.81
	DACE-H	54.36	76.73	**63.40**
3	Zero-shot SigLIP	**62.60**	26.41	37.15
	DACE-S	53.07	78.55	63.34
	DACE-M	44.87	**81.57**	57.89
	DACE-H	56.70	74.37	**64.34**
4	Zero-shot SigLIP	**84.36**	40.03	54.30
	DACE-S	65.01	**83.96**	73.28
	DACE-M	70.88	80.87	**75.55**
	DACE-H	70.76	80.02	75.11
5	Zero-shot SigLIP	**77.90**	33.86	47.16
	DACE-S	52.77	85.77	64.23
	DACE-M	56.29	**86.29**	66.57
	DACE-H	58.86	82.20	**68.01**

Table 12: SigLIP image-to-text retrieval on DomainNet, averaged over the same available forget-domain configurations for all methods at each $|\mathcal{D}_f|$.

$\|\mathcal{D}_f\|$	Method	Image-to-Text Retrieval								
		Mem↑			*For*↑			*H*↑		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1	Zero-shot SigLIP	**62.47**	**82.29**	**86.15**	37.53	17.71	13.85	44.24	23.30	18.39
	DACE-S	41.68	65.35	72.28	88.03	80.17	76.21	55.51	71.66	73.96
	DACE-M	35.48	60.41	68.25	**95.64**	**89.62**	**85.44**	51.38	71.76	75.46
	DACE-H	43.45	69.20	76.30	91.13	84.93	82.29	**58.44**	**76.15**	**79.08**
2	Zero-shot SigLIP	**62.47**	**82.29**	**86.15**	37.53	17.71	13.85	46.36	26.83	21.58
	DACE-S	39.13	63.30	70.65	84.88	74.28	68.87	52.90	67.91	69.39
	DACE-M	32.77	57.11	64.99	**89.72**	**80.36**	**74.84**	47.30	66.11	69.12
	DACE-H	43.95	68.42	75.41	85.60	75.44	70.66	**57.64**	**71.54**	**72.79**
3	Zero-shot SigLIP	**52.60**	**72.29**	**77.00**	27.66	7.71	4.70	36.25	13.93	8.86
	DACE-S	31.05	54.54	62.30	83.66	69.97	63.21	45.29	61.30	62.75
	DACE-M	27.52	50.39	58.85	**94.10**	**85.44**	**79.85**	42.58	63.40	**67.76**
	DACE-H	39.53	62.49	69.45	82.67	70.45	63.50	**53.49**	**66.23**	66.34
4	Zero-shot SigLIP	**73.50**	**95.87**	**98.36**	43.04	24.50	19.95	54.29	39.03	33.18
	DACE-S	38.50	70.21	79.48	85.20	74.84	69.86	53.04	72.45	74.36
	DACE-M	45.40	75.86	83.38	**89.56**	**80.19**	**74.31**	60.25	**77.96**	**78.58**
	DACE-H	48.47	79.27	87.56	83.31	72.68	68.29	**61.29**	75.83	76.73
5	Zero-shot SigLIP	**72.80**	**93.72**	**96.53**	39.60	20.00	15.92	51.25	32.96	27.34
	DACE-S	42.85	70.32	78.76	91.67	84.45	80.42	58.30	76.59	79.43
	DACE-M	43.85	71.69	80.20	**95.30**	**89.03**	**84.83**	59.62	78.90	81.94
	DACE-H	48.20	76.95	84.75	92.39	86.46	83.26	**63.17**	**81.13**	**83.72**

Table 13: SigLIP text-to-image retrieval on DomainNet, averaged over the same available forget-domain configurations for all methods at each $|\mathcal{D}_f|$.

$\|\mathcal{D}_f\|$	Method	Text-to-Image Retrieval								
		Mem↑			*For*↑			*H*↑		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1	Zero-shot SigLIP	**64.58**	**83.58**	**87.34**	35.42	16.42	12.66	42.92	21.99	17.15
	DACE-S	43.79	71.25	78.77	91.37	86.50	83.57	**58.20**	**77.85**	**80.93**
	DACE-M	41.28	63.19	70.09	85.33	76.25	72.09	54.06	67.44	69.46
	DACE-H	32.47	60.99	71.63	**92.10**	**88.27**	**86.73**	46.94	71.71	78.29
2	Zero-shot SigLIP	**64.58**	**83.58**	**87.34**	35.42	16.42	12.66	45.00	25.20	20.00
	DACE-S	37.67	66.77	75.45	89.55	83.33	80.43	52.87	**74.08**	**77.80**
	DACE-M	40.55	62.18	69.10	80.37	66.19	59.44	**53.60**	63.47	63.09
	DACE-H	26.97	56.50	68.03	**89.82**	**83.71**	**81.04**	41.28	66.80	73.47
3	Zero-shot SigLIP	**53.34**	**73.31**	**78.51**	24.17	6.16	3.84	33.27	11.36	7.32
	DACE-S	37.22	61.17	68.73	85.64	78.51	74.79	**51.89**	**68.76**	**71.63**
	DACE-M	37.07	55.94	62.73	75.41	60.39	54.06	49.70	58.08	58.07
	DACE-H	22.85	45.53	56.58	**87.17**	**81.09**	**78.81**	36.21	58.32	65.87
4	Zero-shot SigLIP	**74.47**	**95.46**	**98.40**	40.36	22.36	18.20	52.35	36.24	30.72
	DACE-S	42.24	75.21	84.49	88.33	**83.53**	**81.51**	57.15	**79.15**	**82.97**
	DACE-M	49.24	74.24	81.48	79.66	64.89	58.28	**60.86**	69.25	67.96
	DACE-H	27.69	59.39	75.20	**89.22**	83.15	80.80	42.27	69.29	77.90
5	Zero-shot SigLIP	**75.28**	**94.49**	**97.06**	37.56	18.60	14.61	50.10	31.09	25.40
	DACE-S	53.17	83.24	88.61	92.60	87.13	83.66	**67.43**	**85.00**	**85.89**
	DACE-M	41.16	67.28	75.09	88.20	76.54	70.24	55.96	71.02	71.57
	DACE-H	38.44	73.36	83.64	**93.21**	**89.19**	**87.01**	54.30	80.43	85.22

K Robustness to Few-Shot Relearning Attacks

We evaluate whether information about the forgotten Clipart domain can be recovered from the representations of the unlearned model using a small set of fresh Clipart examples. The attack support images are disjoint from both the samples used to perform unlearning and the held-out evaluation set. Throughout these experiments, the complete CLIP/DACE encoder remains frozen.

Frozen linear probe. We first train a linear classifier on top of the frozen image representations. Given an image x, we extract its normalized representation $\mathbf{z}(x) = \frac{f_I(x)}{\|f_I(x)\|_2}$, where f_I denotes the unlearned image encoder. Only the parameters of the new linear classifier are optimized using $k \in \{1, 2, 4, 8\}$ fresh Clipart images per class. Consequently, this attack measures whether class-discriminative information remains decodable from the frozen representation; it does not modify or relearn the released CLIP/DACE model.

Prototypical Relearning Attack. Following the Prototypical Relearning Attack (PRA) of Ha et al. (2026), we construct an image prototype for each class:

$$\mathbf{p}_c = \operatorname{norm}\left(\frac{1}{k}\sum_{i=1}^{k}\operatorname{norm}\big(f_I(x_i^c)\big)\right). \tag{61}$$

We adapt PRA to CLIP by interpolating each prototype with the corresponding class-text representation $\mathbf{w}_c$ of the unlearned model:

$$\widetilde{\mathbf{w}}_c(\alpha) = \operatorname{norm}\left((1-\alpha)\mathbf{w}_c + \alpha\mathbf{p}_c\right), \qquad \hat{y}(x) = \arg\max_c \mathbf{z}(x)^\top \widetilde{\mathbf{w}}_c(\alpha). \tag{62}$$

Here, $\alpha = 0$ preserves the unlearned zero-shot classifier, whereas $\alpha = 1$ completely replaces its class-text representations with forget-domain image prototypes. No encoder parameters are updated. We regard an attack configuration as utility preserving when its macro-averaged accuracy over the five retained domains decreases by at most one percentage point.

Few-shot recovery results. Table 14 reports the forgotten-domain accuracy obtained with different support-set sizes. The linear probe becomes increasingly effective as the number of examples grows. At eight shots per class, it reaches 64.27%, 20.81%, and 29.07% Clipart accuracy for DACE-S, DACE-M, and DACE-H, respectively. The corresponding unlearned zero-shot accuracies are 46.41%, 8.76%, and 14.16%. This result indicates that some Clipart class information remains decodable from the frozen image representations. However, because the probe is trained only using Clipart examples, its output layer is specialized to the forget domain and does not preserve performance on the retained domains. It should therefore be interpreted as a representation-level information probe rather than a utility-preserving recovery of the original model.

Table 14: Few-shot frozen-feature attacks on the Clipart-unlearned models. Shots refer only to the attack support set; evaluation uses the same held-out test set. Original CLIP obtains 69.67% Clipart accuracy, and the encoder remains frozen.

Method	Shots/class	Unlearned ZS	PRA	Linear Probe
DACE-S	1	46.41	46.40	29.58
	2	46.41	46.40	44.42
	4	46.41	46.40	56.01
	8	46.41	46.40	64.27
DACE-M	1	8.76	8.78	6.60
	2	8.76	8.78	10.12
	4	8.76	8.78	15.73
	8	8.76	8.78	20.81
DACE-H	1	14.16	14.17	7.54
	2	14.16	14.17	12.80
	4	14.16	14.17	20.76
	8	14.16	14.17	29.07

PRA sensitivity and retained utility. Table 15 examines the effect of the PRA interpolation coefficient using eight support images per class. Under the one-percentage-point retained-utility con-

straint, $\alpha = 0$ is selected for all three DACE variants. For DACE-S, $\alpha = 0.2$ increases Clipart accuracy only modestly, from 46.40% to 49.79%, while reducing retained-domain accuracy by 33.52 percentage points. Full prototype replacement ($\alpha = 1$) produces an even larger retained-domain drop of 48.26 points. For DACE-M and DACE-H, the evaluated nonzero interpolation coefficients both decrease Clipart accuracy and severely degrade retained-domain performance.

Consequently, the evaluated PRA does not provide utility-preserving recovery of the forgotten domain. This conclusion is deliberately attack-specific: it does not establish that forgotten information is impossible to recover under every possible adversary. Rather, it shows that, across the evaluated PRA coefficients, any nontrivial prototype intervention either fails to improve forget-domain accuracy or causes a substantial collapse in retained-domain utility.

Table 15: PRA interpolation sensitivity with eight fresh Clipart images per class. We show the selected utility-preserving setting ($\alpha = 0$), the smallest tested nonzero interpolation ($\alpha = 0.2$), and full prototype replacement ($\alpha = 1$). Retain accuracy is macro-averaged over the five retained domains; feasible indicates a drop of at most one percentage point.

Method	α	Clipart Acc.	Retain Acc.	Retain Drop	Feasible
DACE-S	0.00	46.40	52.70	−0.01	✓
	0.20	49.79	19.17	33.52	–
	1.00	34.65	4.44	48.26	–
DACE-M	0.00	8.78	44.94	−0.02	✓
	0.20	5.97	6.96	37.95	–
	1.00	5.86	0.90	44.01	–
DACE-H	0.00	14.17	48.38	−0.02	✓
	0.20	13.81	8.40	39.96	–
	1.00	7.20	1.17	47.18	–

L Limitations and Future Work

Our study addresses approximate alignment unlearning rather than certified removal of training-data influence. The margin guarantees in Appendix A concern prompt-based classification and top-K retrieval; they do not establish deletion of all information about forgotten samples from the model parameters or visual representation. Indeed, the relearning analysis shows that some class-discriminative information can remain accessible to a new probe. Our use of "task-agnostic" therefore refers to editing the shared image–text similarity geometry rather than a task-specific head; the formal guarantees in this paper are restricted to prompt-based classification and cross-modal retrieval; generation, reasoning, and other VLM capabilities are not covered by either the theory or the present experiments. DomainNet retrieval additionally relies on BLIP-2-generated captions rather than native image–text pairs, so future work should evaluate domain-conditional alignment unlearning on datasets with naturally occurring multimodal supervision. Further directions include validating the top-K margin objective suggested by our analysis, stronger adversarial/relearning evaluations, native-caption benchmarks, larger VLMs, and guarantees that connect behavioral alignment suppression to stronger notions of information removal.

M Per-domain Results

Tables 16–18 provide the analogous detailed results for DomainNet, complementing the summary shown in Tables 1, 4, and 5 and Tables 19–24 expand the summary in Table 3 by reporting the complete per-domain results for each experimental setting on Office-Home and Mini DomainNet. These appendix tables provide a more fine-grained breakdown of the comparisons in the main paper and show how the observed retain-forget trade-offs behave across individual domain choices.

Table 16: Per-domain results on DomainNet for $|\mathcal{D}_f| = 1$. Gray shading denotes the forget-domain columns. Within each column, **bold** indicates the best result and underlined indicates the second-best result. For forget-domain columns, lower is better; otherwise, higher is better.

$\lvert\mathcal{D}_f\rvert$	$\mathcal{D}_f$	Task	Method	Per-domain results (%)																	
				clipart			*infograph*			*painting*			*quickdraw*			*real*			*sketch*		
				R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10
		Classification	Zero-shot CLIP	69.67			48.81			**64.27**			14.42			**83.62**			**64.10**		
			ADU (Kawamura et al., 2025)	33.41			**50.61**			60.61			**18.37**			80.11			61.08		
			CLIPErase (Yang et al., 2025)	10.28			2.83			39.80			1.70			58.23			34.09		
			DACE-S (ours)	44.60			47.15			58.39			15.72			81.20			60.06		
			DACE-M (ours)	**7.10**			43.59			52.02			12.57			65.22			49.91		
			DACE-H (ours)	11.88			44.93			54.65			13.62			67.83			52.75		
	{*clipart*}	Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	**67.90**	**93.59**	**97.68**	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	**2.03**	5.87	8.91	0.22	0.51	0.87	0.88	1.75	2.27	0.51	2.54	3.41	0.94	2.10	2.97	2.75	6.09	8.19
			CLIPErase (Yang et al., 2025)	5.96	9.70	10.83	16.42	30.15	36.69	27.06	53.56	63.25	3.62	12.28	18.62	29.17	56.16	65.47	18.15	39.25	48.82
			DACE-S (ours)	28.83	50.90	58.65	49.28	73.34	79.95	62.57	88.96	93.63	8.01	22.86	32.06	64.89	92.72	96.99	46.45	80.94	89.60
			DACE-M (ours)	8.38	17.23	20.67	46.02	69.44	76.83	53.90	81.45	88.28	7.14	21.34	30.69	51.12	79.06	86.99	35.87	66.13	77.41
			DACE-H (ours)	4.06	**5.51**	**8.41**	**71.59**	**86.38**	**91.30**	**77.03**	**96.22**	**97.67**	**24.93**	**50.73**	**64.06**	62.61	92.46	96.23	**77.39**	**92.75**	**95.65**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	**95.00**	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	**2.61**	6.81	**8.62**	0.36	0.65	0.94	1.39	2.12	3.14	0.07	0.36	0.65	2.97	4.20	5.43	4.42	7.32	9.35
			CLIPErase (Yang et al., 2025)	5.20	9.62	11.38	21.41	37.83	44.80	24.05	51.83	61.40	1.92	7.90	12.90	26.74	54.42	64.96	17.10	38.52	48.93
			DACE-S (ours)	26.89	47.49	54.88	47.19	70.07	76.86	61.67	88.28	92.76	7.17	19.86	29.31	64.13	92.79	96.38	42.27	77.26	86.18
			DACE-M (ours)	3.33	7.65	10.06	39.66	62.47	69.74	49.57	77.42	84.36	7.07	18.80	26.96	45.62	74.64	81.99	30.92	62.60	72.14
			DACE-H (ours)	3.77	**6.67**	10.14	**66.09**	**81.74**	**88.70**	**77.33**	**93.89**	**96.80**	**21.74**	**47.25**	**58.55**	**76.52**	91.88	94.20	**64.35**	**84.64**	**89.56**
		Classification	Zero-shot CLIP	69.67			48.81			**64.27**			14.42			**83.62**			64.10		
			ADU (Kawamura et al., 2025)	**72.97**			9.64			63.38			**32.10**			79.64			**65.94**		
			CLIPErase (Yang et al., 2025)	43.07			**4.33**			34.19			0.29			61.81			15.64		
			DACE-S (ours)	69.30			13.44			61.36			17.10			81.74			62.57		
			DACE-M (ours)	63.45			5.66			58.27			15.94			74.78			58.24		
			DACE-H (ours)	60.00			6.09			52.91			14.20			71.59			55.94		
	{*infograph*}	Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.03	5.87	8.91	**0.22**	**0.51**	**0.87**	0.88	1.75	2.27	0.51	2.54	3.41	0.94	2.10	2.97	2.75	6.09	8.19
			CLIPErase (Yang et al., 2025)	19.91	43.91	54.48	2.86	5.03	5.54	11.57	22.20	26.61	3.04	9.46	14.49	27.39	53.15	63.04	17.10	39.87	49.76
			DACE-S (ours)	**54.34**	86.31	94.36	12.71	16.56	17.74	65.40	91.90	95.67	8.66	23.44	33.37	68.55	93.88	97.36	49.11	82.03	91.12
			DACE-M (ours)	50.60	82.00	90.67	8.34	12.01	13.99	60.46	88.65	94.27	8.19	23.01	31.96	60.65	89.49	94.06	43.98	78.50	86.76
			DACE-H (ours)	38.55	**98.26**	**99.13**	2.03	3.19	4.06	**87.21**	**98.26**	**99.71**	**24.64**	**53.33**	**64.93**	**89.86**	**99.42**	**99.71**	**84.93**	**96.81**	**97.68**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	2.61	6.81	8.62	**0.36**	**0.65**	**0.94**	1.39	2.12	3.14	0.07	0.36	0.65	2.97	4.20	5.43	4.42	7.32	9.35
			CLIPErase (Yang et al., 2025)	19.25	44.57	56.24	2.72	4.74	5.47	15.38	34.79	45.80	0.36	1.88	3.62	26.81	55.36	65.29	16.08	37.14	46.67
			DACE-S (ours)	54.34	86.79	93.52	14.73	20.31	22.00	64.53	90.80	94.95	8.59	22.79	31.41	70.29	94.71	97.83	48.45	82.69	90.00
			DACE-M (ours)	44.42	78.92	87.71	2.39	5.29	7.09	56.01	85.26	91.29	7.14	20.80	29.06	53.48	87.54	92.64	39.72	74.50	83.78
			DACE-H (ours)	**61.74**	**95.36**	**97.10**	0.58	1.45	3.19	**84.01**	**97.09**	**98.26**	**16.52**	**44.64**	**58.84**	**88.99**	**96.52**	97.68	**78.26**	**94.20**	**95.94**
		Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			**83.62**			**64.10**		
			ADU (Kawamura et al., 2025)	**72.90**			**53.30**			24.27			**27.46**			78.12			63.62		
			CLIPErase (Yang et al., 2025)	43.43			9.77			**6.07**			0.33			27.54			19.13		
			DACE-S (ours)	69.30			46.93			25.56			17.28			81.05			61.00		
			DACE-M (ours)	63.52			46.68			11.38			15.76			70.40			56.38		
			DACE-H (ours)	58.84			42.03			9.88			16.23			67.25			54.20		
	{*painting*}	Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.17	6.30	8.12	0.72	1.38	2.03	**0.73**	**1.39**	**2.05**	0.51	1.45	2.83	1.16	3.62	4.78	1.52	3.33	5.00
			CLIPErase (Yang et al., 2025)	30.04	56.75	66.41	18.84	34.74	41.76	7.01	10.59	11.80	5.18	16.16	24.35	21.34	41.12	49.20	22.59	45.98	54.53
			DACE-S (ours)	54.92	86.53	93.96	51.97	75.32	81.64	25.97	34.98	36.52	7.97	23.26	32.54	67.36	93.51	97.43	46.89	81.19	90.18
			DACE-M (ours)	42.85	78.48	88.33	42.27	67.72	75.03	11.87	19.15	22.01	7.43	22.07	30.69	46.38	76.92	86.59	38.96	71.95	81.92
			DACE-H (ours)	**84.93**	**97.39**	**98.84**	**72.75**	**86.38**	**89.28**	2.33	5.81	6.39	**25.80**	**51.88**	**62.61**	**88.99**	**98.84**	**100.00**	**79.71**	**94.78**	**97.39**
1		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	3.84	6.23	7.03	0.94	1.23	1.59	**0.36**	**0.88**	**1.46**	0.07	0.51	1.09	2.54	4.06	4.64	2.54	4.71	5.72
			CLIPErase (Yang et al., 2025)	30.04	58.14	67.87	25.56	47.34	55.56	6.97	11.08	12.51	1.92	7.39	12.72	28.33	53.95	63.66	20.99	49.18	58.24
			DACE-S (ours)	54.12	86.35	93.45	47.48	71.65	77.89	26.91	37.43	40.18	7.83	20.80	30.29	65.94	94.17	97.43	45.14	79.67	88.07
			DACE-M (ours)	43.58	77.53	86.06	38.49	63.02	70.58	4.64	9.57	12.33	6.85	19.42	28.33	46.38	75.14	82.75	34.88	69.08	79.70
			DACE-H (ours)	**77.97**	**94.49**	**97.97**	**65.80**	**82.32**	**87.54**	3.49	6.11	7.85	**19.71**	**44.64**	**57.68**	**80.87**	**97.39**	98.26	**75.94**	**91.30**	**93.91**
		Classification	Zero-shot CLIP	69.67			48.81			**64.27**			14.42			**83.62**			64.10		
			ADU (Kawamura et al., 2025)	**74.64**			**53.01**			63.09			1.30			81.59			**64.56**		
			CLIPErase (Yang et al., 2025)	47.60			5.36			44.82			0.33			64.75			18.92		
			DACE-S (ours)	70.22			49.21			62.34			0.43			82.10			62.68		
			DACE-M (ours)	67.87			46.20			57.52			0.11			79.78			60.49		
			DACE-H (ours)	69.56			47.83			61.05			**0.00**			79.13			62.61		
	{*quickdraw*}	Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.97	7.10	9.56	0.14	0.87	1.38	0.66	1.53	1.97	0.07	0.36	1.01	1.09	2.03	3.04	2.03	6.59	8.62
			CLIPErase (Yang et al., 2025)	34.83	64.00	73.77	17.48	29.49	34.01	32.53	59.97	69.84	0.04	0.22	0.36	46.59	77.43	86.20	15.39	29.94	35.18
			DACE-S (ours)	56.53	88.47	95.06	54.57	76.61	82.52	66.00	91.93	96.19	0.04	0.18	0.43	70.69	95.25	98.55	50.24	83.85	92.11
			DACE-M (ours)	57.01	87.30	94.18	51.93	75.43	81.64	64.53	92.05	96.04	0.14	0.36	0.47	68.95	94.82	98.48	48.53	82.65	90.47
			DACE-H (ours)	**89.86**	**98.84**	**99.42**	**76.81**	**88.70**	**92.17**	**90.70**	**98.84**	**99.71**	**0.00**	**0.00**	**0.29**	**94.20**	**100.00**	**100.00**	**90.14**	**98.26**	**99.42**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	3.84	7.10	8.12	0.29	0.72	1.09	0.44	1.02	1.61	**0.00**	0.36	0.65	1.67	2.32	2.75	3.12	6.52	7.97
			CLIPErase (Yang et al., 2025)	35.02	64.00	73.07	19.21	33.68	40.18	40.45	66.68	74.48	0.04	**0.18**	**0.36**	44.93	77.90	85.22	14.08	28.37	36.38
			DACE-S (ours)	54.26	87.56	94.15	50.68	73.56	79.98	66.68	90.73	95.25	0.04	**0.18**	**0.36**	69.78	95.40	98.22	48.27	81.99	90.07
			DACE-M (ours)	52.51	85.51	92.54	44.03	68.97	76.06	60.16	88.13	93.74	0.04	0.29	0.43	65.62	93.48	97.39	44.56	79.88	88.29
			DACE-H (ours)	**87.25**	**98.55**	**99.13**	**74.49**	**87.54**	**91.30**	**90.70**	**99.13**	**99.13**	**0.00**	0.29	0.58	**94.20**	**99.13**	**100.00**	**88.41**	**97.10**	**97.97**
		Classification	Zero-shot CLIP	69.67			48.81			**64.27**			14.42			83.62			64.10		
			ADU (Kawamura et al., 2025)	**71.52**			**51.78**			62.13			**30.94**			34.78			**65.87**		
			CLIPErase (Yang et al., 2025)	22.06			11.16			16.25			14.06			8.33			35.10		
			DACE-S (ours)	68.35			46.97			58.69			14.49			69.71			60.97		
			DACE-M (ours)	55.91			42.20			47.30			15.80			10.58			56.46		
			DACE-H (ours)	55.36			41.16			46.51			15.94			**7.83**			57.39		
	{*real*}	Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	**90.50**	**95.33**	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.61	6.96	9.35	0.43	0.72	0.94	0.51	1.39	1.83	0.36	1.59	3.33	**0.65**	**1.30**	**1.81**	1.52	4.57	6.01
			CLIPErase (Yang et al., 2025)	22.58	45.41	54.19	26.33	48.59	56.85	9.16	18.66	22.84	5.00	16.52	24.53	5.51	8.48	9.53	22.92	51.36	62.24
			DACE-S (ours)	51.01	83.13	91.18	51.01	73.30	80.50	58.95	86.66	92.39	8.08	23.26	32.61	32.83	60.51	72.32	47.76	80.94	89.81
			DACE-M (ours)	39.37	70.80	80.24	42.93	67.83	75.65	43.12	70.11	78.89	7.57	21.92	31.85	10.00	19.02	24.31	39.87	72.32	82.65
			DACE-H (ours)	**77.39**	**92.46**	**96.23**	**69.56**	**84.93**	**88.41**	**70.93**	89.24	93.02	**24.35**	**52.46**	**65.22**	10.44	13.91	18.55	**79.71**	**94.78**	**97.68**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	**92.65**	**96.57**	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	4.64	7.83	9.71	0.14	0.51	0.80	0.58	1.02	1.83	0.07	0.51	0.80	**0.65**	**1.23**	**1.81**	4.13	7.39	9.28
			CLIPErase (Yang et al., 2025)	17.71	40.03	49.65	18.47	35.84	43.74	18.09	38.41	47.87	4.28	13.30	21.92	4.82	8.41	9.82	18.55	44.38	55.33
			DACE-S (ours)	51.92	85.84	92.64	48.40	72.90	78.41	59.59	87.75	93.44	7.54	21.88	30.65	44.17	75.29	84.53	47.25	80.21	88.54
			DACE-M (ours)	38.86	69.27	77.46	38.52	62.84	71.06	41.46	68.15	75.80	6.56	20.00	28.73	5.91	10.76	13.30	36.38	70.17	80.61
			DACE-H (ours)	**75.65**	**91.88**	**94.78**	**63.77**	**82.32**	**87.54**	**72.09**	86.92	89.83	**19.13**	**48.41**	**63.77**	8.41	15.94	20.58	**74.78**	**91.01**	**94.49**
		Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			**83.62**			64.10		
			ADU (Kawamura et al., 2025)	**70.51**			**53.30**			**64.33**			**29.35**			82.10			21.74		
			CLIPErase (Yang et al., 2025)	11.56			7.71			30.04			0.29			54.49			**2.91**		
			DACE-S (ours)	69.96			47.89			61.25			16.77			82.17			42.02		
			DACE-M (ours)	61.69			45.24			56.31			13.95			77.86			7.42		
			DACE-H (ours)	54.49			44.64			49.71			16.52			72.17			4.06		
	{*sketch*}	Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.10	6.01	8.12	0.29	0.51	0.94	0.58	1.17	1.75	0.22	1.38	2.46	0.51	0.87	1.74	**0.94**	**1.74**	**2.03**
			CLIPErase (Yang et al., 2025)	23.16	45.88	54.63	28.61	48.66	56.41	18.81	34.53	40.63	0.11	0.25	0.58	35.94	64.53	73.19	3.42	5.67	6.04
			DACE-S (ours)	54.55	84.81	**93.16**	50.97	73.67	80.54	62.87	89.41	94.61	7.90	21.92	31.52	68.19	94.31	98.19	25.10	43.03	49.29
			DACE-M (ours)	44.16	75.05	82.99	47.63	71.21	77.56	54.73	82.40	88.81	6.81	20.91	30.40	61.34	89.71	95.87	8.73	13.90	16.30
			DACE-H (ours)	**70.14**	**88.12**	91.88	**65.51**	**86.09**	**89.28**	**84.30**	**96.22**	**97.67**	**22.32**	**51.30**	**62.03**	**79.42**	**97.97**	**99.13**	2.90	6.96	8.99
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	**92.65**	**96.57**	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	2.97	4.71	5.14	0.36	0.65	0.80	0.51	1.10	1.53	0.00	0.58	0.87	1.09	1.59	1.96	**0.29**	**1.16**	**1.67**
			CLIPErase (Yang et al., 2025)	20.60	43.21	52.29	22.84	40.95	48.62	29.14	53.03	61.63	0.07	0.25	0.54	40.04	70.65	78.70	3.35	5.57	5.93
			DACE-S (ours)	52.87	85.18	**92.13**	47.92	71.47	77.97	59.74	87.90	93.67	6.56	20.04	29.06	66.12	94.17	97.72	24.59	48.20	58.46
			DACE-M (ours)	43.25	73.66	82.58	42.60	66.88	73.67	53.26	81.12	87.22	5.98	19.17	27.83	61.99	90.07	94.93	5.06	9.53	11.60
			DACE-H (ours)	**68.41**	**85.80**	87.83	**64.64**	**79.13**	**84.64**	**77.03**	92.44	95.35	**16.81**	**42.03**	**58.84**	**87.83**	**97.97**	**98.55**	5.80	8.99	11.59

Table 17: Per-domain results on DomainNet for $|\mathcal{D}_f| \in \{2, 3, 4\}$. Gray shading denotes the forget-domain columns. Within each column, **bold** indicates the best result and underlined indicates the second-best result. For forget-domain columns, lower is better; otherwise, higher is better.

$\lvert\mathcal{D}_f\rvert$	$\mathcal{D}_f$	Task	Method	Per-domain results (%)																	
				clipart			*infograph*			*painting*			*quickdraw*			*real*			*sketch*		
				R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10
2	{*clipart*, *infograph*}	Classification	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	**0.87**	**2.10**	**2.68**	**0.29**	**0.58**	**0.94**	0.80	1.53	1.97	0.58	1.88	2.83	0.87	2.25	3.04	2.68	7.83	10.07
			CLIPErase (Yang et al., 2025)	2.49	4.54	5.53	0.92	1.69	1.91	7.16	13.49	16.47	1.45	5.47	9.09	11.96	24.67	30.33	5.75	15.17	21.54
			DACE-S (ours)	28.47	48.92	54.77	13.18	18.07	19.32	60.53	87.98	93.93	8.41	23.19	33.12	64.75	91.92	96.81	45.44	79.63	88.58
			DACE-M (ours)	13.17	28.58	35.78	8.15	13.22	16.09	55.07	84.06	90.31	7.61	21.88	31.30	51.41	83.37	90.36	36.85	71.81	82.61
			DACE-H (ours)	1.45	18.55	37.68	2.03	3.77	4.93	**81.98**	**97.97**	**98.84**	**24.06**	**54.49**	**66.09**	**84.93**	**98.84**	**99.13**	**80.87**	**96.81**	**98.84**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	**0.94**	**1.67**	**1.96**	0.22	**0.65**	**1.23**	3.36	7.75	9.43	0.14	0.36	1.01	5.72	8.55	9.64	5.14	8.99	11.59
			CLIPErase (Yang et al., 2025)	1.79	4.06	5.05	0.88	1.06	1.43	12.85	25.71	31.59	1.20	4.82	7.75	11.45	22.14	26.77	5.82	15.46	20.84
			DACE-S (ours)	30.30	57.41	66.23	10.65	15.76	17.70	60.16	87.03	92.42	7.79	21.77	31.52	63.88	93.19	97.03	43.98	78.39	86.76
			DACE-M (ours)	6.88	15.84	19.47	2.90	6.13	7.67	51.68	79.61	87.15	6.05	18.51	27.17	44.89	78.73	86.34	35.21	67.70	77.81
			DACE-H (ours)	1.45	44.93	57.97	**0.00**	1.16	2.32	**83.43**	**95.64**	**97.09**	**20.29**	**48.12**	**60.29**	**84.93**	**96.81**	**98.55**	**69.56**	**88.99**	**94.49**
	{*painting*, *quickdraw*}	Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			**83.62**			**64.10**		
			ADU (Kawamura et al., 2025)	**73.91**			**53.88**			21.27			3.26			77.97			64.06		
			CLIPErase (Yang et al., 2025)	18.70			7.71			**3.24**			0.29			12.97			8.04		
			DACE-S (ours)	69.63			47.30			40.14			1.30			82.61			60.79		
			DACE-M (ours)	65.61			47.74			17.72			**0.25**			75.04			58.38		
			DACE-H (ours)	46.67			38.55			4.07			0.29			47.83			42.32		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.39	6.38	8.91	0.29	0.87	1.45	**0.36**	**0.95**	**1.46**	0.14	0.43	1.09	0.94	2.17	3.04	2.17	3.99	4.86
			CLIPErase (Yang et al., 2025)	19.47	39.96	48.19	25.08	44.91	52.33	4.60	7.73	9.20	**0.07**	**0.18**	**0.36**	17.72	31.49	37.39	10.44	21.06	24.48
			DACE-S (ours)	56.02	86.83	94.25	53.18	75.32	81.45	34.72	48.28	51.19	0.25	0.47	0.72	68.30	93.91	98.30	47.40	80.61	89.89
			DACE-M (ours)	47.46	81.19	90.71	48.07	72.60	78.52	15.45	26.46	31.36	**0.07**	0.40	0.87	53.08	84.35	92.14	40.81	74.75	85.81
			DACE-H (ours)	**86.67**	**97.97**	**98.84**	**75.07**	**85.51**	**90.44**	51.45	77.62	86.34	0.29	2.03	3.77	**92.17**	**99.42**	**99.71**	**81.45**	**96.52**	**97.97**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	4.86	9.49	11.16	0.29	1.09	1.30	**0.36**	**0.66**	**1.32**	0.22	0.36	1.16	1.59	2.32	2.61	2.46	5.43	6.52
			CLIPErase (Yang et al., 2025)	23.97	46.03	54.63	27.91	47.85	55.38	5.43	8.86	9.99	**0.04**	**0.18**	**0.36**	20.22	37.79	44.60	10.48	21.64	27.57
			DACE-S (ours)	52.03	85.36	92.68	49.03	71.83	78.08	34.38	50.66	55.15	**0.04**	0.29	0.43	66.96	94.42	97.90	44.34	79.41	87.63
			DACE-M (ours)	45.01	79.03	87.74	41.79	67.31	75.50	9.99	18.28	21.48	**0.04**	**0.18**	0.47	51.30	83.41	90.11	37.98	72.97	83.56
			DACE-H (ours)	**80.87**	**96.23**	**98.26**	**68.41**	**86.09**	**89.86**	25.58	50.00	59.88	1.45	3.48	4.64	**85.51**	**97.39**	98.26	**70.72**	**91.59**	**95.94**
	{*real*, *sketch*}	Classification	Zero-shot CLIP	**69.67**			48.81			**64.27**			14.42			83.62			64.10		
			ADU (Kawamura et al., 2025)	68.84			**52.36**			61.84			**30.73**			35.22			19.27		
			CLIPErase (Yang et al., 2025)	0.66			0.73			0.56			0.25			**0.47**			**0.47**		
			DACE-S (ours)	68.46			47.74			57.52			15.18			64.67			37.29		
			DACE-M (ours)	54.70			42.20			50.06			16.45			13.73			8.04		
			DACE-H (ours)	44.93			39.71			40.70			13.04			9.28			6.38		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	2.32	6.09	8.99	0.29	0.87	1.09	0.58	1.10	1.24	0.58	1.52	2.68	**0.58**	**1.16**	**1.67**	1.01	**1.59**	**2.25**
			CLIPErase (Yang et al., 2025)	2.78	9.15	13.54	5.91	16.53	22.48	0.87	2.64	3.43	0.04	0.22	0.54	1.88	3.48	4.28	**0.62**	1.96	2.55
			DACE-S (ours)	50.57	82.11	90.93	49.25	73.23	80.32	57.67	85.45	91.86	7.54	23.04	32.06	40.22	67.86	78.12	23.75	40.20	46.89
			DACE-M (ours)	36.74	67.14	77.83	41.57	68.82	76.20	43.72	72.11	82.40	6.81	21.20	29.38	15.11	31.23	38.91	10.26	22.08	27.57
			DACE-H (ours)	**64.93**	**86.09**	**92.75**	**58.55**	**82.32**	**88.70**	**73.84**	**91.57**	**95.64**	**21.16**	**50.73**	**61.45**	12.75	33.33	50.15	4.93	10.14	13.04
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	**92.65**	**96.57**	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	4.06	6.74	8.26	0.51	0.72	0.94	0.22	0.95	1.61	0.00	0.14	0.22	**0.36**	**0.87**	**1.30**	0.58	1.81	2.17
			CLIPErase (Yang et al., 2025)	1.39	4.43	7.25	4.37	11.64	17.74	2.11	5.54	7.12	0.04	0.22	0.51	0.87	2.14	2.94	**0.55**	**1.24**	**1.56**
			DACE-S (ours)	52.10	**84.19**	**91.29**	45.61	69.26	76.64	58.50	86.02	92.08	6.70	20.51	29.20	46.74	76.09	83.80	22.99	43.94	53.26
			DACE-M (ours)	39.37	67.87	77.53	38.82	63.94	71.25	44.85	71.81	79.46	6.05	18.80	28.01	10.00	19.38	24.06	6.84	13.79	18.01
			DACE-H (ours)	**62.03**	82.03	86.67	**60.29**	**78.26**	**85.22**	**67.73**	82.27	87.50	**18.84**	**44.35**	**56.52**	10.14	19.71	28.99	9.86	16.81	22.03
3	{*clipart*, *infograph*, *painting*}	Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			**83.62**			**64.10**		
			ADU (Kawamura et al., 2025)	26.01			6.74			20.76			**34.71**			75.51			62.10		
			CLIPErase (Yang et al., 2025)	**0.40**			**0.26**			**0.38**			0.29			0.33			0.33		
			DACE-S (ours)	48.85			14.76			34.00			17.75			80.04			60.02		
			DACE-M (ours)	21.33			6.10			14.96			14.96			70.51			55.04		
			DACE-H (ours)	24.35			5.80			12.79			14.20			65.51			54.20		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.65	1.23	2.10	0.22	0.65	1.16	0.44	1.46	1.90	0.65	1.81	2.54	0.58	1.96	2.54	2.03	6.30	9.13
			CLIPErase (Yang et al., 2025)	**0.07**	**0.37**	**0.51**	**0.04**	**0.18**	**0.40**	**0.11**	**0.34**	**0.49**	0.04	0.25	0.29	0.04	0.29	0.69	0.04	0.33	0.66
			DACE-S (ours)	30.26	51.30	58.80	14.91	21.04	23.50	30.27	44.89	49.68	8.19	22.79	33.19	64.28	92.17	96.59	43.40	77.96	87.85
			DACE-M (ours)	11.74	27.11	34.87	8.56	15.13	17.92	14.02	24.69	28.57	8.55	23.70	33.33	48.19	78.88	88.51	36.45	70.72	81.19
			DACE-H (ours)	19.13	53.33	69.56	2.90	4.93	6.96	33.43	63.08	74.13	**25.80**	**53.04**	**66.38**	**86.96**	**97.97**	**98.55**	**79.42**	**93.62**	**96.81**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.80	1.38	1.96	0.22	0.87	1.59	0.73	1.32	1.75	0.58	1.01	1.16	2.97	4.20	5.07	3.41	6.67	9.20
			CLIPErase (Yang et al., 2025)	**0.07**	**0.33**	**0.77**	**0.04**	**0.22**	**0.51**	**0.07**	**0.34**	**0.60**	0.07	0.40	0.65	0.11	0.36	0.51	0.11	0.40	0.73
			DACE-S (ours)	32.34	61.07	70.73	13.99	22.33	25.71	32.34	52.43	61.10	7.17	20.18	29.82	63.44	91.85	96.67	41.32	75.45	84.39
			DACE-M (ours)	9.15	20.56	25.87	4.33	8.26	10.36	9.05	17.57	20.96	7.75	21.41	30.54	46.16	78.44	86.63	34.38	67.48	78.28
			DACE-H (ours)	1.74	42.03	55.36	0.87	2.61	5.51	26.16	47.97	58.14	**17.97**	**47.54**	**61.74**	**82.90**	**96.23**	97.10	**70.72**	**89.56**	**91.30**
4	{*clipart*, *infograph*, *painting*, *quickdraw*}	Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			**83.62**			**64.10**		
			ADU (Kawamura et al., 2025)	26.59			8.19			19.96			2.61			75.22			62.25		
			CLIPErase (Yang et al., 2025)	**0.26**			**0.33**			**0.30**			**0.36**			0.22			0.44		
			DACE-S (ours)	58.51			23.03			47.34			1.96			80.72			57.18		
			DACE-M (ours)	23.34			7.57			16.74			1.52			70.14			51.47		
			DACE-H (ours)	26.67			7.83			17.44			0.87			68.41			55.36		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.58	1.74	2.39	0.36	0.94	1.59	0.66	1.46	1.90	0.07	0.51	0.94	1.09	2.10	3.12	2.25	5.87	8.77
			CLIPErase (Yang et al., 2025)	**0.04**	**0.22**	**0.51**	**0.00**	**0.07**	**0.26**	**0.04**	**0.23**	**0.45**	**0.04**	**0.22**	**0.36**	0.04	0.43	0.51	0.04	0.26	0.58
			DACE-S (ours)	40.07	71.90	82.55	21.92	37.68	45.35	36.52	59.52	68.94	1.05	3.37	5.72	62.46	92.25	97.39	42.41	78.68	88.11
			DACE-M (ours)	16.50	39.22	50.02	11.64	21.37	25.82	17.68	33.40	40.33	2.10	5.80	7.97	46.63	81.52	90.36	36.59	72.06	83.41
			DACE-H (ours)	14.49	64.64	75.94	20.58	27.83	32.17	48.26	70.93	80.52	0.29	3.77	5.51	**88.99**	**98.55**	**99.71**	**78.84**	**95.65**	**97.97**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.51	1.16	1.67	0.22	0.58	1.01	0.44	0.95	1.17	**0.00**	0.36	0.65	2.46	3.19	3.84	3.91	8.04	10.14
			CLIPErase (Yang et al., 2025)	**0.07**	**0.44**	**0.77**	**0.11**	**0.26**	**0.44**	**0.04**	**0.26**	**0.30**	0.04	**0.18**	**0.43**	0.07	0.25	0.40	0.04	0.29	0.44
			DACE-S (ours)	33.30	61.98	73.18	13.81	22.04	25.49	35.66	57.75	66.87	0.04	0.22	0.76	62.39	92.14	96.92	41.40	75.01	83.30
			DACE-M (ours)	10.90	26.75	33.81	4.66	9.07	11.71	11.27	22.01	26.84	0.65	2.10	3.48	44.46	77.86	86.52	34.09	68.06	78.47
			DACE-H (ours)	2.32	52.46	64.35	0.58	2.32	2.90	24.13	45.06	57.85	1.45	5.80	8.41	**85.22**	**95.94**	97.68	**69.56**	**91.30**	**96.52**

Table 18: Per-domain results on DomainNet for $|\mathcal{D}_f| = 5$. Gray shading denotes the forget-domain columns. Within each column, **bold** indicates the best result and underlined indicates the second-best result. For forget-domain columns, lower is better; otherwise, higher is better.

$\lvert\mathcal{D}_f\rvert$	$\mathcal{D}_f$	Task	Method	Per-domain results (%)																	
				clipart			*infograph*			*painting*			*quickdraw*			*real*			*sketch*		
				R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10	R@1/Acc.	R@5	R@10
5	{*clipart, infograph, painting, quickdraw, real*}	Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			83.62			**64.10**		
			ADU (Kawamura et al., 2025)	11.52			2.83			10.82			3.19			9.71			59.93		
			CLIPErase (Yang et al., 2025)	**0.33**			**0.29**			**0.26**			0.33			**0.07**			0.40		
			DACE-S (ours)	34.65			9.62			30.49			2.10			43.77			58.31		
			DACE-M (ours)	11.53			2.79			5.69			0.33			6.45			52.20		
			DACE-H (ours)	24.35			3.77			13.37			**0.29**			21.74			56.81		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.72	1.52	2.46	0.36	0.58	0.94	0.58	1.17	1.39	0.07	0.36	0.72	0.14	0.80	1.52	2.10	5.94	7.83
			CLIPErase (Yang et al., 2025)	**0.00**	**0.11**	**0.22**	**0.04**	**0.18**	**0.33**	**0.04**	**0.19**	**0.30**	**0.04**	**0.18**	**0.29**	**0.00**	**0.11**	**0.33**	0.00	0.11	0.22
			DACE-S (ours)	24.73	52.43	63.78	19.39	37.35	45.02	28.12	49.87	60.12	0.98	2.21	3.30	36.85	71.23	83.33	43.40	78.83	87.49
			DACE-M (ours)	9.84	22.76	29.02	5.69	12.71	16.49	8.41	16.81	21.48	0.43	1.23	1.96	7.57	19.24	25.73	38.16	71.52	82.21
			DACE-H (ours)	33.33	56.81	64.93	12.75	23.77	28.41	25.58	43.31	52.03	0.87	2.90	5.80	44.93	67.25	75.07	**83.48**	**96.23**	**97.68**
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.87	1.52	2.03	0.14	0.72	1.09	0.51	0.88	1.32	0.29	0.43	0.65	0.14	0.72	1.09	4.78	10.51	12.17
			CLIPErase (Yang et al., 2025)	**0.00**	**0.00**	**0.11**	**0.00**	**0.11**	**0.33**	**0.04**	**0.19**	**0.30**	**0.04**	**0.11**	**0.29**	**0.07**	**0.18**	**0.29**	0.04	0.22	0.44
			DACE-S (ours)	18.70	34.58	41.16	4.63	8.67	10.83	18.24	33.24	40.22	0.14	0.58	0.76	22.36	41.85	48.44	42.82	78.36	86.32
			DACE-M (ours)	4.65	12.11	16.10	1.76	3.56	5.07	3.47	7.50	10.74	**0.04**	0.43	0.69	2.68	6.85	9.67	32.88	67.26	78.03
			DACE-H (ours)	3.77	41.45	50.44	1.16	2.61	3.77	11.63	25.58	33.14	2.90	7.25	9.56	17.97	35.65	42.90	**75.36**	**93.91**	**96.23**
	{*clipart, infograph, painting, quickdraw, sketch*}	Classification	Zero-shot CLIP	69.67			48.81			64.27			14.42			**83.62**			64.10		
			ADU (Kawamura et al., 2025)	12.32			8.19			16.23			0.80			74.42			6.59		
			CLIPErase (Yang et al., 2025)	**0.33**			**0.26**			**0.23**			**0.29**			0.29			**0.29**		
			DACE-S (ours)	29.53			9.95			17.04			1.16			75.07			12.15		
			DACE-M (ours)	20.71			6.65			14.85			0.69			69.93			8.37		
			DACE-H (ours)	18.55			5.80			13.08			0.87			68.12			5.51		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.36	0.58	1.16	0.22	0.72	1.38	0.22	0.88	1.46	0.07	0.36	0.72	0.65	1.81	2.75	0.14	0.80	1.38
			CLIPErase (Yang et al., 2025)	**0.00**	**0.11**	**0.37**	**0.04**	**0.22**	**0.40**	**0.07**	**0.15**	**0.34**	**0.04**	**0.18**	**0.40**	0.07	0.14	0.36	**0.04**	**0.15**	**0.33**
			DACE-S (ours)	28.80	58.07	68.64	22.36	39.33	46.02	34.75	58.20	68.68	1.96	6.41	10.18	61.45	91.52	96.49	22.04	43.03	53.04
			DACE-M (ours)	13.06	32.27	40.14	8.74	16.23	19.98	13.87	26.27	32.19	1.27	3.33	4.93	51.16	81.59	89.75	6.77	16.26	21.57
			DACE-H (ours)	33.62	55.65	62.61	18.84	30.14	33.33	44.77	65.12	71.51	0.29	2.03	4.06	**86.96**	**98.84**	**99.71**	22.90	42.32	49.56
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	**98.33**	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.36	0.65	1.01	0.72	0.94	1.38	1.02	1.68	2.27	**0.00**	0.36	0.65	6.38	10.29	12.25	0.51	1.23	1.59
			CLIPErase (Yang et al., 2025)	**0.04**	**0.22**	**0.37**	**0.00**	**0.26**	**0.40**	**0.04**	**0.15**	**0.34**	0.04	**0.18**	**0.33**	0.00	0.22	0.33	**0.04**	**0.26**	**0.40**
			DACE-S (ours)	14.38	26.82	35.78	10.80	17.19	19.72	19.79	31.93	38.15	0.07	0.33	0.87	56.16	88.80	94.78	9.13	19.10	26.01
			DACE-M (ours)	6.48	15.88	21.70	3.38	6.68	9.55	7.05	16.32	20.81	0.22	1.05	1.38	48.44	81.20	88.44	3.17	7.75	11.20
			DACE-H (ours)	1.45	41.74	52.17	0.87	2.61	6.09	20.35	40.70	50.29	0.58	2.32	4.64	**86.67**	**97.10**	97.97	6.38	23.48	31.59
	{*clipart, infograph, quickdraw, real, sketch*}	Classification	Zero-shot CLIP	69.67			48.81			**64.27**			14.42			83.62			64.10		
			ADU (Kawamura et al., 2025)	8.33			4.57			57.02			0.65			21.67			8.77		
			CLIPErase (Yang et al., 2025)	**0.33**			**0.26**			0.30			0.29			**0.36**			**0.33**		
			DACE-S (ours)	34.06			17.85			52.13			0.29			68.73			24.84		
			DACE-M (ours)	6.99			3.12			47.46			**0.25**			20.18			5.28		
			DACE-H (ours)	0.87			0.29			32.27			0.29			0.87			1.16		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.29	1.16	1.67	**0.00**	0.36	0.72	1.02	2.27	3.51	0.07	0.36	0.72	0.22	0.72	1.52	0.36	0.80	1.30
			CLIPErase (Yang et al., 2025)	**0.04**	**0.11**	**0.33**	0.04	**0.15**	**0.29**	0.00	0.11	0.34	**0.04**	**0.22**	**0.43**	**0.07**	**0.25**	**0.47**	**0.07**	**0.18**	**0.33**
			DACE-S (ours)	28.54	58.03	68.39	28.46	48.29	57.69	57.44	86.05	92.50	2.43	6.45	9.67	54.17	87.86	95.00	17.90	35.90	45.40
			DACE-M (ours)	11.42	28.87	37.47	8.78	16.75	21.41	49.04	79.83	87.18	1.20	3.33	4.96	19.42	44.38	54.06	7.97	18.37	25.46
			DACE-H (ours)	12.17	26.09	42.90	31.88	48.12	56.52	**80.23**	**94.77**	**97.97**	0.87	2.03	4.35	54.20	91.01	95.07	12.75	30.14	43.77
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	**92.65**	**96.57**	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.22	0.58	1.09	0.22	0.36	0.65	1.61	2.92	3.87	**0.00**	0.29	0.72	0.22	0.72	1.16	0.36	0.80	1.30
			CLIPErase (Yang et al., 2025)	**0.00**	**0.15**	**0.40**	**0.00**	**0.18**	**0.44**	0.00	0.11	0.38	0.04	**0.18**	**0.36**	**0.04**	**0.18**	**0.36**	**0.04**	**0.15**	**0.33**
			DACE-S (ours)	6.70	18.00	25.28	10.50	19.50	25.16	55.18	83.83	89.63	0.07	**0.18**	0.62	34.93	64.28	76.96	7.02	18.48	24.92
			DACE-M (ours)	3.55	8.45	12.33	1.65	4.00	5.47	49.30	77.53	83.72	**0.00**	0.25	0.72	9.31	19.60	25.51	2.87	7.35	9.89
			DACE-H (ours)	7.83	22.61	32.46	5.22	16.52	26.09	**72.38**	90.70	95.35	0.29	2.03	3.48	23.77	53.04	68.70	4.35	18.55	25.22
	{*clipart, painting, quickdraw, real, sketch*}	Classification	Zero-shot CLIP	69.67			**48.81**			64.27			14.42			83.62			64.10		
			ADU (Kawamura et al., 2025)	8.41			47.14			7.53			0.72			18.12			6.09		
			CLIPErase (Yang et al., 2025)	**0.37**			0.26			**0.11**			**0.29**			**0.29**			**0.29**		
			DACE-S (ours)	36.99			47.67			19.90			1.30			61.92			21.28		
			DACE-M (ours)	5.34			42.16			3.96			**0.29**			11.96			2.98		
			DACE-H (ours)	6.96			41.16			3.49			0.58			14.49			3.77		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.29	1.16	1.67	0.00	0.36	0.72	1.02	2.27	3.51	0.07	0.36	0.72	0.22	0.72	1.52	0.36	0.80	1.30
			CLIPErase (Yang et al., 2025)	**0.07**	**0.26**	**0.40**	0.04	0.22	0.37	**0.00**	**0.11**	**0.38**	**0.04**	**0.14**	**0.36**	**0.07**	**0.22**	**0.40**	**0.07**	**0.22**	**0.44**
			DACE-S (ours)	31.61	61.65	74.20	49.10	73.56	80.50	26.91	43.88	51.68	0.62	1.96	3.01	50.04	82.61	92.03	18.59	36.67	44.56
			DACE-M (ours)	9.48	22.94	30.66	42.86	67.65	76.53	8.26	19.94	26.69	0.83	2.54	3.73	15.54	37.79	49.89	5.38	13.31	18.01
			DACE-H (ours)	8.12	20.58	34.20	**68.41**	**82.61**	**86.96**	28.49	45.64	56.69	0.29	1.74	3.77	25.51	63.48	73.62	8.70	28.41	39.71
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.22	0.58	1.09	0.22	0.36	0.65	1.61	2.92	3.87	**0.00**	0.29	0.72	0.22	0.72	1.16	0.36	0.80	1.30
			CLIPErase (Yang et al., 2025)	**0.04**	**0.18**	**0.29**	0.07	0.18	0.44	**0.04**	**0.23**	**0.45**	0.07	**0.18**	**0.36**	**0.04**	**0.18**	**0.36**	**0.04**	**0.18**	**0.36**
			DACE-S (ours)	19.32	38.05	48.34	45.28	71.10	77.45	16.66	29.78	35.39	0.07	0.47	0.98	40.47	71.12	80.22	11.28	26.23	33.25
			DACE-M (ours)	3.62	9.40	12.04	39.26	63.24	70.58	2.64	5.73	8.29	0.04	0.22	0.43	5.51	12.43	17.32	1.64	4.76	7.02
			DACE-H (ours)	1.74	14.49	21.45	**58.55**	**79.13**	**84.35**	4.65	13.95	20.06	0.29	1.45	4.35	9.56	25.22	37.39	1.45	6.96	10.14
	{*infograph, painting, quickdraw, real, sketch*}	Classification	Zero-shot CLIP	**69.67**			48.81			64.27			14.42			83.62			64.10		
			ADU (Kawamura et al., 2025)	67.39			3.48			9.36			1.52			18.62			13.19		
			CLIPErase (Yang et al., 2025)	0.40			**0.37**			**0.19**			**0.25**			**0.29**			**0.22**		
			DACE-S (ours)	62.97			21.52			29.48			1.05			62.46			42.67		
			DACE-M (ours)	61.60			6.24			10.76			0.60			23.24			18.44		
			DACE-H (ours)	60.58			3.48			10.17			1.16			21.45			26.38		
		Text-to-Image Retrieval	Zero-shot CLIP	50.53	82.18	90.96	53.62	75.91	81.56	64.57	90.50	95.33	7.79	21.30	30.87	67.90	93.59	97.68	44.42	78.50	87.27
			ADU (Kawamura et al., 2025)	0.29	1.16	1.67	**0.00**	0.36	0.72	1.02	2.27	3.51	0.07	0.36	0.72	0.22	0.72	1.52	0.36	0.80	1.30
			CLIPErase (Yang et al., 2025)	0.00	0.18	0.40	0.04	**0.18**	**0.40**	**0.00**	**0.11**	**0.41**	**0.04**	**0.18**	**0.36**	**0.04**	**0.14**	**0.40**	**0.00**	**0.22**	**0.36**
			DACE-S (ours)	48.52	82.66	90.93	23.76	44.29	52.81	34.04	62.76	73.84	2.17	7.14	10.54	42.17	76.20	87.14	29.28	61.73	73.55
			DACE-M (ours)	38.20	71.90	84.19	7.01	16.34	20.82	7.99	20.13	26.39	0.87	2.43	3.59	12.17	33.84	44.02	10.99	27.86	35.72
			DACE-H (ours)	**77.10**	**96.52**	**97.97**	10.14	15.94	21.45	30.23	48.84	56.40	0.58	2.32	3.19	39.71	68.70	77.10	36.81	57.10	66.96
		Image-to-Text Retrieval	Zero-shot CLIP	45.37	80.94	89.32	49.43	71.76	78.26	67.02	92.65	96.57	6.99	18.66	26.16	69.28	95.00	98.33	48.27	80.68	88.72
			ADU (Kawamura et al., 2025)	0.22	0.58	1.09	0.22	0.36	0.65	1.61	2.92	3.87	**0.00**	0.29	0.72	0.22	0.72	1.16	0.36	0.80	1.30
			CLIPErase (Yang et al., 2025)	0.00	0.22	0.48	**0.11**	**0.18**	**0.33**	**0.04**	**0.15**	**0.34**	0.04	**0.18**	**0.36**	**0.04**	**0.22**	**0.36**	**0.00**	**0.26**	**0.47**
			DACE-S (ours)	46.21	82.25	89.75	13.77	25.56	31.69	19.75	36.30	46.55	0.22	1.56	3.08	34.38	64.42	74.13	22.84	49.36	60.60
			DACE-M (ours)	42.96	76.84	85.91	3.19	7.01	8.96	3.39	10.74	14.40	0.07	0.47	0.69	7.97	17.93	23.99	6.47	15.82	20.23
			DACE-H (ours)	**48.12**	**89.86**	**94.78**	0.29	1.74	4.06	11.05	20.93	29.36	0.87	3.19	6.38	18.84	36.52	45.80	17.39	34.49	44.93

Table 19: Per-domain results on Office-Home for $|\mathcal{D}_f| = 1$, reporting classification accuracy for each domain, where gray shading denotes the forget domain column. Within each column of a given forget-set block, **bold** indicates the best result and underlined indicates the second-best result, with lower accuracy preferred for forget domain columns and higher accuracy preferred otherwise.

$\mathcal{D}_f$	Method	Per-domain accuracy (%)				Unlearning metrics (%)		
		art	*clipart*	*product*	*real*	*Mem*↑	*For*↑	*H*↑
{*art*}	Zero-shot CLIP	83.40	68.86	90.43	**87.53**	82.39	16.60	27.64
	ADU (Kawamura et al., 2025)	43.65	70.76	90.78	79.00	80.28	56.35	66.05
	DACE-S (ours)	34.21	**75.22**	**92.55**	80.64	**82.89**	65.79	73.22
	DACE-M (ours)	23.77	69.23	89.65	71.68	76.95	76.23	76.58
	DACE-H (ours)	**13.08**	70.76	90.07	67.10	76.06	**86.92**	**81.09**
{*clipart*}	Zero-shot CLIP	83.40	68.86	90.43	87.53	87.75	31.14	45.97
	ADU (Kawamura et al., 2025)	80.38	19.52	89.79	85.66	86.11	80.48	83.17
	DACE-S (ours)	84.40	7.60	**92.34**	87.03	88.53	92.40	90.40
	DACE-M (ours)	84.40	5.56	91.56	**88.60**	88.83	94.44	91.54
	DACE-H (ours)	**85.16**	**2.85**	92.13	87.67	**88.86**	**97.15**	**92.82**
{*product*}	Zero-shot CLIP	83.40	68.86	90.43	**87.53**	**79.43**	9.57	17.09
	ADU (Kawamura et al., 2025)	81.76	70.47	34.97	81.72	77.40	65.03	70.68
	DACE-S (ours)	84.15	**74.05**	25.18	81.43	79.20	74.82	76.84
	DACE-M (ours)	79.25	66.30	17.30	74.91	72.57	82.69	77.28
	DACE-H (ours)	**85.79**	73.25	**8.01**	76.27	77.23	**91.99**	**83.96**
{*real*}	Zero-shot CLIP	**83.40**	68.86	**90.43**	87.53	**80.61**	12.47	21.60
	ADU (Kawamura et al., 2025)	71.82	70.39	85.18	54.77	76.55	45.23	56.81
	DACE-S (ours)	74.72	**74.12**	88.51	61.15	79.93	38.85	52.27
	DACE-M (ours)	60.63	64.03	76.17	40.00	68.07	60.00	63.45
	DACE-H (ours)	60.00	70.47	76.45	**25.09**	70.50	**74.91**	**72.60**

Table 20: Per-domain results on Office-Home for $|\mathcal{D}_f| = 2$, reporting classification accuracy for each domain, where gray shading denotes the forget domain column. Within each column of a given forget-set block, **bold** indicates the best result and underlined indicates the second-best result, with lower accuracy preferred for forget domain columns and higher accuracy preferred otherwise.

$\mathcal{D}_f$	Method	Per-domain accuracy (%)				Unlearning metrics (%)		
		art	*clipart*	*product*	*real*	*Mem*↑	*For*↑	*H*↑
{*art*, *clipart*}	Zero-shot CLIP	83.40	68.86	90.43	**87.53**	**88.98**	25.80	40.00
	ADU (Kawamura et al., 2025)	40.75	16.74	90.28	80.93	85.63	74.43	79.60
	DACE-S (ours)	29.69	5.48	**92.20**	80.43	86.35	85.62	85.94
	DACE-M (ours)	30.31	6.65	91.35	78.35	84.88	84.65	84.76
	DACE-H (ours)	**17.36**	**2.63**	91.21	74.12	82.71	**91.96**	**87.07**
{*art*, *product*}	Zero-shot CLIP	83.40	68.86	90.43	**87.53**	**78.28**	12.11	20.97
	ADU (Kawamura et al., 2025)	44.15	70.47	42.06	76.06	73.29	57.19	64.06
	DACE-S (ours)	36.86	**73.03**	30.85	69.75	71.37	66.98	68.37
	DACE-M (ours)	31.19	66.59	21.49	60.07	63.30	75.01	68.40
	DACE-H (ours)	**21.26**	71.13	**10.71**	61.08	66.05	**85.49**	**74.47**
{*art*, *real*}	Zero-shot CLIP	83.40	68.86	**90.43**	87.53	79.81	13.97	23.78
	ADU (Kawamura et al., 2025)	16.48	68.06	85.96	30.82	77.14	74.38	75.74
	DACE-S (ours)	12.58	**72.81**	87.09	25.16	**80.06**	79.41	79.70
	DACE-M (ours)	8.05	70.39	84.54	19.79	77.57	84.48	80.86
	DACE-H (ours)	**3.90**	69.88	81.70	**11.90**	75.88	**91.00**	**82.74**
{*clipart*, *product*}	Zero-shot CLIP	**83.40**	68.86	90.43	**87.53**	**86.03**	20.19	32.71
	ADU (Kawamura et al., 2025)	83.27	22.37	39.79	82.37	82.69	68.79	75.10
	DACE-S (ours)	77.86	11.26	32.77	79.78	79.09	77.83	78.44
	DACE-M (ours)	74.34	10.82	25.04	76.63	75.80	81.97	78.73
	DACE-H (ours)	79.75	**3.95**	**8.30**	75.27	76.89	**93.84**	**84.53**
{*clipart*, *real*}	Zero-shot CLIP	**83.40**	68.86	**90.43**	87.53	**87.89**	21.72	34.83
	ADU (Kawamura et al., 2025)	70.82	16.74	87.09	52.19	81.22	65.36	72.43
	DACE-S (ours)	76.60	15.72	88.30	67.02	84.08	58.38	68.81
	DACE-M (ours)	56.35	8.19	72.20	37.56	66.49	76.98	69.04
	DACE-H (ours)	62.26	**3.36**	86.88	**33.12**	78.00	**81.61**	**79.77**
{*product*, *real*}	Zero-shot CLIP	**83.40**	68.86	90.43	87.53	**74.20**	11.02	19.18
	ADU (Kawamura et al., 2025)	69.94	69.37	18.65	40.50	69.58	70.48	70.02
	DACE-S (ours)	71.82	**74.49**	16.45	38.78	73.51	72.44	72.91
	DACE-M (ours)	65.28	71.42	12.06	30.75	69.16	78.65	73.60
	DACE-H (ours)	65.16	72.22	**3.33**	**21.51**	69.63	**87.63**	**77.55**

Table 21: Per-domain results on Office-Home for $|\mathcal{D}_f| = 3$, reporting classification accuracy for each domain, where gray shading denotes the forget domain column. Within each column of a given forget-set block, **bold** indicates the best result and underlined indicates the second-best result, with lower accuracy preferred for forget domain columns and higher accuracy preferred otherwise.

$\mathcal{D}_f$	Method	Per-domain accuracy (%)				Unlearning metrics (%)		
		art	*clipart*	*product*	*real*	*Mem*↑	*For*↑	*H*↑
{*art*, *clipart*, *product*}	Zero-shot CLIP	83.40	68.86	90.43	**87.53**	**87.53**	19.39	31.75
	ADU (Kawamura et al., 2025)	47.67	16.89	44.18	77.56	77.56	65.49	70.97
	DACE-S (ours)	38.99	6.94	37.94	74.62	74.62	73.69	**74.15**
	DACE-M (ours)	32.33	10.38	27.09	64.87	64.87	78.14	70.69
	DACE-H (ours)	**26.79**	**3.73**	**21.99**	64.95	64.95	**83.94**	73.13
{*art*, *clipart*, *real*}	Zero-shot CLIP	83.40	68.86	**90.43**	87.53	**90.43**	20.57	33.52
	ADU (Kawamura et al., 2025)	9.94	7.31	86.52	27.96	86.52	84.01	85.23
	DACE-S (ours)	7.92	4.61	88.44	24.44	88.44	86.88	87.64
	DACE-M (ours)	7.42	6.29	79.65	22.65	79.65	87.04	82.85
	DACE-H (ours)	**2.39**	**1.68**	87.16	**15.27**	87.16	**92.83**	**89.90**
{*art*, *product*, *real*}	Zero-shot CLIP	83.40	68.86	90.43	87.53	68.86	12.25	20.80
	ADU (Kawamura et al., 2025)	11.57	60.45	16.95	14.55	60.45	85.17	70.59
	DACE-S (ours)	9.18	**73.76**	6.81	6.59	**73.76**	92.75	**82.16**
	DACE-M (ours)	8.80	65.42	9.79	10.39	65.42	90.19	75.83
	DACE-H (ours)	**6.04**	71.27	**3.76**	**3.01**	71.27	**96.03**	81.79
{*clipart*, *product*, *real*}	Zero-shot CLIP	**83.40**	68.86	90.43	87.53	**83.40**	17.61	29.08
	ADU (Kawamura et al., 2025)	69.94	14.99	19.08	**37.49**	69.94	76.11	72.87
	DACE-S (ours)	71.95	8.85	**11.49**	38.57	71.95	**80.33**	**75.78**
	DACE-M (ours)	75.09	7.67	24.47	44.95	75.09	74.19	74.64
	DACE-H (ours)	71.70	**6.94**	11.84	42.44	71.70	79.54	75.35

Table 22: Per-domain results on Mini-DomainNet for $|\mathcal{D}_f| = 1$, reporting classification accuracy for each domain, where gray shading denotes the forget domain column. Within each column of a given forget-set block, **bold** indicates the best result and underlined indicates the second-best result, with lower accuracy preferred for forget domain columns and higher accuracy preferred otherwise.

$\mathcal{D}_f$	Method	Per-domain accuracy (%)				Unlearning metrics (%)		
		clipart	*painting*	*real*	*sketch*	*Mem*↑	*For*↑	*H*↑
{*clipart*}	Zero-shot CLIP	83.02	79.21	**90.48**	**78.41**	**82.70**	16.98	28.18
	ADU (Kawamura et al., 2025)	28.94	79.21	85.08	75.19	79.82	71.06	75.17
	DACE-S (ours)	15.08	78.84	85.87	76.72	80.48	84.92	82.63
	DACE-M (ours)	13.76	**79.63**	84.60	75.71	79.98	86.24	**82.98**
	DACE-H (ours)	**8.89**	74.23	82.91	70.16	75.77	**91.11**	82.61
{*painting*}	Zero-shot CLIP	83.02	79.21	**90.48**	78.41	83.97	20.79	33.33
	ADU (Kawamura et al., 2025)	83.86	28.52	87.35	77.51	82.91	71.48	76.75
	DACE-S (ours)	86.35	16.03	87.72	**79.63**	**84.57**	83.97	84.25
	DACE-M (ours)	**87.04**	14.07	86.61	78.57	84.07	85.93	84.99
	DACE-H (ours)	84.60	**9.79**	83.92	77.25	81.92	**90.21**	**85.87**
{*real*}	Zero-shot CLIP	**83.02**	**79.21**	90.48	**78.41**	**80.21**	9.52	17.03
	ADU (Kawamura et al., 2025)	82.91	76.46	32.12	77.67	79.01	67.88	73.00
	DACE-S (ours)	82.65	75.56	19.42	77.30	78.50	80.58	79.50
	DACE-M (ours)	81.64	73.39	13.97	77.14	77.39	86.03	**81.47**
	DACE-H (ours)	75.77	64.87	**9.37**	73.33	71.32	**90.63**	79.81
{*sketch*}	Zero-shot CLIP	83.02	79.21	**90.48**	78.41	**84.23**	21.59	34.37
	ADU (Kawamura et al., 2025)	81.90	79.84	88.25	20.48	83.33	79.52	81.37
	DACE-S (ours)	**83.44**	**80.74**	88.52	8.41	**84.23**	91.59	87.74
	DACE-M (ours)	82.49	78.47	88.94	8.41	83.30	91.59	87.25
	DACE-H (ours)	80.90	79.10	88.41	**5.98**	82.80	**94.02**	**88.05**

Table 23: Per-domain results on Mini-DomainNet for $|\mathcal{D}_f| = 2$, reporting classification accuracy for each domain, where gray shading denotes the forget domain column. Within each column of a given forget-set block, **bold** indicates the best result and underlined indicates the second-best result, with lower accuracy preferred for forget domain columns and higher accuracy preferred otherwise.

$\mathcal{D}_f$	Method	Per-domain accuracy (%)				Unlearning metrics (%)		
		clipart	*painting*	*real*	*sketch*	*Mem*↑	*For*↑	*H*↑
{*clipart*, *painting*}	Zero-shot CLIP	83.02	79.21	**90.48**	**78.41**	**84.44**	18.89	30.87
	ADU (Kawamura et al., 2025)	28.52	24.39	83.65	74.39	79.02	73.55	76.17
	DACE-S (ours)	14.97	14.50	83.44	75.29	79.36	85.26	82.21
	DACE-M (ours)	15.34	15.50	82.06	74.81	78.44	84.58	81.39
	DACE-H (ours)	**9.58**	**9.89**	79.58	72.65	76.11	**90.26**	**82.58**
{*clipart*, *real*}	Zero-shot CLIP	83.02	**79.21**	90.48	**78.41**	**78.81**	13.25	22.69
	ADU (Kawamura et al., 2025)	20.85	75.93	27.94	74.60	75.26	75.61	75.41
	DACE-S (ours)	11.90	73.76	14.18	73.23	73.49	86.96	**79.62**
	DACE-M (ours)	15.19	72.06	14.97	74.60	73.33	84.92	78.68
	DACE-H (ours)	**7.20**	68.36	**7.04**	70.00	69.18	**92.88**	79.15
{*clipart*, *sketch*}	Zero-shot CLIP	83.02	**79.21**	**90.48**	78.41	**84.84**	19.29	31.43
	ADU (Kawamura et al., 2025)	24.87	77.04	84.29	13.23	80.66	80.95	80.75
	DACE-S (ours)	8.84	77.14	84.71	4.44	80.93	93.36	**86.70**
	DACE-M (ours)	11.64	76.35	84.23	6.30	80.29	91.03	85.32
	DACE-H (ours)	**7.88**	76.19	82.75	**4.02**	79.47	**94.05**	86.12
{*painting*, *real*}	Zero-shot CLIP	83.02	79.21	90.48	**78.41**	80.71	15.16	25.52
	ADU (Kawamura et al., 2025)	80.79	18.78	22.22	75.34	78.07	79.50	78.75
	DACE-S (ours)	**83.49**	9.10	12.65	78.04	**80.77**	89.13	**84.74**
	DACE-M (ours)	82.06	8.78	9.68	75.82	78.94	90.77	84.44
	DACE-H (ours)	79.26	**6.24**	**7.83**	73.97	76.61	**92.96**	83.95
{*painting*, *sketch*}	Zero-shot CLIP	**83.02**	79.21	**90.48**	78.41	**86.75**	21.19	34.06
	ADU (Kawamura et al., 2025)	80.85	24.50	86.03	17.57	83.44	78.97	81.11
	DACE-S (ours)	82.17	14.39	87.73	9.95	84.95	87.83	86.36
	DACE-M (ours)	81.43	16.83	86.14	10.90	83.78	86.14	84.94
	DACE-H (ours)	79.52	**9.84**	84.92	**6.03**	82.22	**92.06**	**86.86**
{*real*, *sketch*}	Zero-shot CLIP	**83.02**	**79.21**	90.48	78.41	**81.11**	15.56	26.11
	ADU (Kawamura et al., 2025)	79.10	75.13	30.69	21.69	77.12	73.81	75.39
	DACE-S (ours)	78.89	75.29	23.28	9.79	77.09	83.47	**80.15**
	DACE-M (ours)	77.67	70.64	21.06	10.21	74.15	84.36	78.91
	DACE-H (ours)	72.49	69.63	**11.32**	**6.72**	71.06	**90.98**	79.77

Table 24: Per-domain results on Mini-DomainNet for $|\mathcal{D}_f| = 3$, reporting classification accuracy for each domain, where gray shading denotes the forget domain column. Within each column of a given forget-set block, **bold** indicates the best result and underlined indicates the second-best result, with lower accuracy preferred for forget domain columns and higher accuracy preferred otherwise.

$\mathcal{D}_f$	Method	Per-domain accuracy (%)				Unlearning metrics (%)		
		clipart	*painting*	*real*	*sketch*	*Mem*↑	*For*↑	*H*↑
{*clipart*, *painting*, *real*}	Zero-shot CLIP	83.02	79.21	90.48	**78.41**	**78.41**	15.77	26.25
	ADU (Kawamura et al., 2025)	13.60	13.49	12.54	72.80	72.80	86.79	79.11
	DACE-S (ours)	11.85	5.34	5.87	76.83	76.83	92.31	**83.85**
	DACE-M (ours)	12.22	6.83	8.47	74.23	74.23	90.83	81.68
	DACE-H (ours)	**5.29**	**2.80**	**2.49**	57.20	57.20	**96.47**	69.73
{*clipart*, *painting*, *sketch*}	Zero-shot CLIP	83.02	79.21	**90.48**	78.41	**90.48**	19.79	32.47
	ADU (Kawamura et al., 2025)	19.52	22.54	83.60	11.01	83.60	82.31	82.89
	DACE-S (ours)	8.73	12.91	84.71	3.44	84.71	91.64	**88.03**
	DACE-M (ours)	11.16	15.40	83.12	5.71	83.12	89.24	86.06
	DACE-H (ours)	**7.88**	**9.42**	81.06	**3.02**	81.06	**93.23**	86.71
{*clipart*, *real*, *sketch*}	Zero-shot CLIP	83.02	**79.21**	90.48	78.41	**79.21**	16.03	26.67
	ADU (Kawamura et al., 2025)	14.87	74.34	25.82	10.16	74.34	83.05	78.40
	DACE-S (ours)	**4.87**	73.12	16.40	**3.17**	73.12	91.85	**81.40**
	DACE-M (ours)	11.06	72.01	24.97	6.08	72.01	85.96	78.33
	DACE-H (ours)	7.46	69.89	**12.22**	4.07	69.89	**92.08**	79.47
{*painting*, *real*, *sketch*}	Zero-shot CLIP	**83.02**	79.21	90.48	78.41	**83.02**	17.30	28.64
	ADU (Kawamura et al., 2025)	78.57	16.19	21.32	20.48	78.57	80.67	79.53
	DACE-S (ours)	82.01	6.61	13.02	11.38	82.01	89.67	**85.66**
	DACE-M (ours)	78.31	10.32	18.57	13.28	78.31	85.94	81.94
	DACE-H (ours)	66.83	**2.91**	**6.72**	**5.24**	66.83	**95.04**	77.82