CSOA: CONCEPT-SELECTIVE ADAPTATION WITH SEMANTIC CONTROL FOR VISION-LANGUAGE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept unlearning in vision-language models (VLMs) is increasingly critical for selective removal of unwanted semantic information. However, practice demonstrates that while existing methods report success on traditional metrics, target recognition may remain recoverable through unconventional queries or visual domains. We identify three critical failure modes: visual bias arises when unlearning on narrow forget sets leaves recognition intact in unseen visual domains; association leakage occurs when suppressing specific text templates leaves alternative expressions as retrieval backdoors; and semantic collapse results when unconstrained adaptation disrupts meaningful post-unlearning predictions and neighboring concepts. To address these challenges, we introduce CSOA, a concept-selective adaptation framework with semantic control. This approach combines a quadrant anchor system specifying the linguistic neighborhood with hierarchical control that guides edited representations toward user-specified semantic ancestors, while preserving non-target utility. Knowledge-matched controls and factorial interventions assess semantic redirection and the retention-focused variant CSOA-R. Tests on three CLIP backbones and 40 additional concepts examine breadth; paired generative comparisons distinguish recognition suppression from answer disclosure.

1 INTRODUCTION

The rapid advancement of vision-language models (VLMs) has necessitated the development of machine unlearning techniques to ensure that models can "forget" specific data points after training (Cao & Yang, 2015; Bourtoule et al., 2021). Beyond individual samples, a more critical challenge is *concept unlearning* (Gandikota et al., 2023), referring to the targeted removal of high-level semantic concepts. Unlike standard data deletion, concept unlearning in multi-modal systems must address persistent cross-modal associations that can survive conventional erasure attempts (Yang et al., 2025).

In the pursuit of robust concept erasure, research has converged on the composition of forget/retain sets and the strategic control of post-unlearning behavior. Regarding data composition, frameworks such as DIET (Kunananthaseelan et al., 2026) explore data-efficient unlearning by using hyperbolic geometry to reduce dependence on retain data. Simultaneously, the focus has shifted toward concept redirection and feature decoupling. For instance, TARF (Zhu et al., 2026) highlights the necessity of decoupling class labels from target concepts when their domains do not perfectly match. While these advancements provide building blocks for concept removal, distribution coverage, linguistic coverage, and semantic redirection still require joint evaluation on the same edited model.

As illustrated in Figure 1, this compartmentalized evaluation leads to three critical failure modes. First, even with carefully curated forget sets, **Visual Bias** persists: when unlearning is confined to a specific visual mode, target recognition can remain in unseen domains. Second, the focus on primary text queries neglects the linguistic neighborhood of VLMs, leading to **Association Leakage**, where aliases and descriptions act as alternative routes to recognition. Finally, improper redirection mechanisms can result in **Semantic Collapse**, where predictions become arbitrary or updates damage neighboring concepts. These failures call for testing both where the edited representation moves and what non-target behavior it preserves.

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

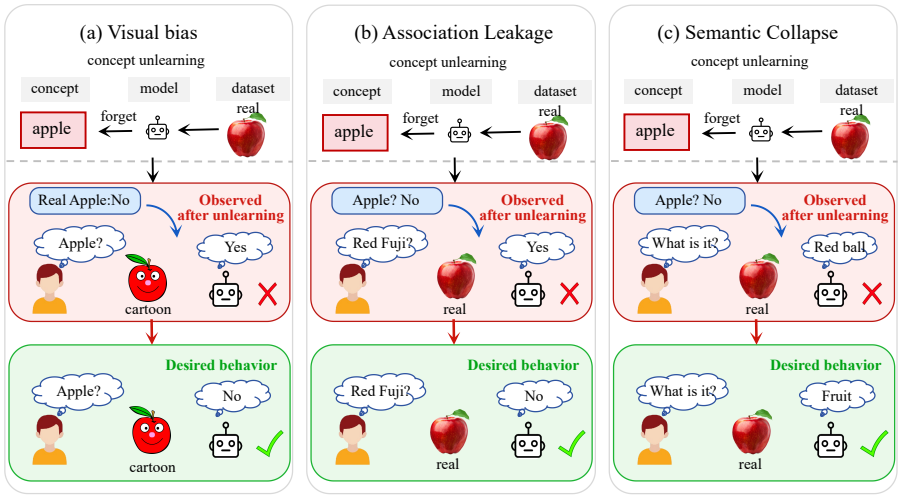

Figure 1: Three structural failure modes of VLM concept unlearning, illustrated for the concept *apple*. (a) Visual bias: recognition survives in an unseen visual style. (b) Association leakage: an alternative expression such as the cultivar name *Red Fuji* recovers the suppressed concept. (c) Semantic collapse: post-unlearning predictions become arbitrary rather than redirected to the intended ancestor. The illustration shows conceptual examples; endpoint and retained-neighborhood measurements test their empirical counterparts.

To bridge these gaps, we propose **CSOA**, a concept-selective adaptation framework with semantic control. The method uses semantic anchors to specify what should be suppressed and preserved, and hierarchical redirection to guide edited concepts toward their semantic ancestors. Its low-rank update balances target suppression with global and neighboring-class utility. Furthermore, we introduce a suite of failure-specific metrics designed to measure cross-domain generalization, linguistic robustness, and semantic structural integrity. Controls with identical semantic resources distinguish the contribution of the adaptation method from that of the knowledge supplied to it.

In summary, we make three contributions:

- We identify and characterize three structural failure modes, namely visual bias, association leakage, and semantic collapse, that conventional forget/retain scores can conceal.
- We introduce a parameter-efficient framework, CSOA, based on low-rank adaptation (Hu et al., 2022), combining a structured semantic specification, hierarchical redirection, and conflict-aware retention. We develop CSOA-R with retention distillation, peer replay, and moderated semantic control to preserve neighboring-class utility, and isolate their effects in a complete factorial study.
- We jointly evaluate visual domains, linguistic expressions, and semantic destinations to expose failures concealed by ordinary forget accuracy. We use knowledge-matched controls, two further CLIP backbones, 40 additional concepts, and paired generative comparisons to test the method's scope.

2 Related Work

Standard machine unlearning. The machine unlearning literature distinguishes exact methods, including retraining-based constructions (Cao & Yang, 2015; Bourtoule et al., 2021), from approximate ones that modify weights post hoc. Fisher Forgetting (Golatkar et al., 2020) uses information-based parameter modification; gradient ascent and SCRUB (Kurmanji et al., 2023) suppress forget behavior with varying retain constraints; SSD (Foster et al., 2024) applies selective synaptic dampening; and SalUn (Fan et al., 2024) restricts updates to saliency-identified parameters. These approaches provide useful forgetting mechanisms, but a fixed dataset loss does not itself test the open-vocabulary linguistic neighborhood or the semantic structure of post-erasure predictions. When applied to VLMs, these additional requirements motivate the three evaluations in Section 1.

VLM concept unlearning. Concept unlearning has been explored in text-to-image diffusion models (Gandikota et al., 2023), while multi-modal encoders introduce distinct cross-modal associations. CLIPErase (Yang et al., 2025) is a close prior work, jointly optimizing forgetting, retention, and cross-modal consistency losses. Synthetic-sample CLIP unlearning (Kravets & Namboodiri, 2025a) addresses data access, while CCUP (Mishra et al., 2025) derives a data-free projection from forget and retain text directions. SLUG (Cai et al., 2025) edits a gradient-selected layer, and Text-image Space adaptation (Kravets & Namboodiri, 2025b) learns a text-driven change to the shared embedding space. Recent approaches also include SineProject (Garg et al., 2026) for vision-language alignment and SAUCE (Geng & Li, 2025) for editing concept-associated sparse features. Multimodal benchmarks further broaden evaluation beyond classification (Liu et al., 2025). MMUnlearner (Huo et al., 2025) updates multimodal language models with a geometry-constrained gradient objective; our LLaVA comparison adapts that objective to the same paired images and prompts, without treating it as a reproduction of its original benchmark. CSOA connects suppression across domains and language queries with an explicit semantic destination. The knowledge-matched and isolated-loss comparisons test this method contribution separately from semantic-resource access.

3 METHOD

3.1 PROBLEM FORMULATION

Let f_θ and g_ϕ be the image and text encoders of a frozen VLM, producing ℓ_2-normalized embeddings $z_{\text{orig}} = f_\theta(x)$ and $t = g_\phi(T)$. For a given image x, we denote its representation after unlearning via adapter Δ as $\hat{z} = f_{\theta+\Delta}(x)$. The cross-modal binding of concept c^* is the set of associated image-text pairs with high joint similarity, $\mathcal{B}(c^*) = \{(x, T) : \langle f_\theta(x), g_\phi(T)\rangle > \eta,\ x \in \mathcal{V}_{c^*},\ T \in \mathcal{S}_{c^*}\}$. Here $\mathcal{V}_{c^*}$ and $\mathcal{S}_{c^*}$ describe the concept's visual distribution and linguistic expressions.

Concept-selective unlearning seeks to suppress the target across its visual and linguistic manifestations while preserving non-target utility. The objectives are: (C1) low target recognition in the source and unseen visual domains; (C2) low residual recognition under alternative expressions; (C3) post-unlearning behavior compatible with a user-specified semantic ancestor; and (Retain) stable predictions for non-target concepts. These objectives are assessed on finite, disjoint evaluation sets; the formal formulation and scope of each measurement are given in Appendix A.1.

3.2 CONCEPT-SELECTIVE ADAPTATION WITH SEMANTIC CONTROL

CSOA combines a structured semantic specification, hierarchical redirection, and conflict-aware retention under a low-rank adaptation objective. The anchor-based concept direction provides an auxiliary regularizer, not a separate source of established gains (Appendix Figure 3). The primary CLIP comparison uses the original source-only recipe with an always-on adapter, keeping checkpoint and strength fixed across diagnostics. An optional router is examined only in Appendix C.6 and is not part of the main method claim.

3.3 QUADRANT ANCHOR SYSTEM (QAS)

To achieve comprehensive concept removal, we first specify the linguistic boundary of the target concept to prevent association leakage. A loss penalizing a single nominated text t_{c^*} = "a photo of a {class}" constrains one direction in the joint embedding space; it does not by itself bound recognition under other expressions. For concept c^* we build four anchor sets in text space: A_{erase} contains target names and descriptions; A_{ctrl} specifies permitted control ancestors; A_{peer} contains related concepts; and A_{reject} contains concepts to protect from interference. Each set is encoded by g_ϕ to produce its embeddings and a normalized centroid $\bar{t}_\bullet$. The raw direction $\Delta_{\text{raw}} = \bar{t}_{\text{erase}} - \bar{t}_{\text{ctrl}}$ approximates the axis between the target and its intended post-unlearning position. The implemented banks are finite: explicit 15-string banks are used for apple and rose, with 15-template banks for the remaining concepts. Independent descriptions and exact-label peers test generalization beyond those training strings (Appendix B.2).

3.4 Anchor-Based Concept Direction

QAS supplies a concept-associated direction for an auxiliary regularizer, without assuming distribution invariance. In the source-only implementation, candidate interference directions consist of rejection, context, peer, and control text directions. Five context anchors are selected using the mean feature of 16 source training targets; when an explicit peer bank is available, its leading two singular vectors provide additional interference directions. Starting from $v_0 = \Delta_{\text{raw}}$, the implementation applies sequential projection,

$$v_{j+1} = v_j - (v_j^\top u_j)u_j, \qquad d_c = \frac{v_J}{\|v_J\|_2}, \tag{1}$$

where each u_j is a normalized candidate direction. This source-only construction does not guarantee simultaneous orthogonality of the candidate directions. Its ordering and auxiliary regularizer are detailed in Appendix B.1.

3.5 Loss Objectives and Training Curriculum

We optimize the adapter weights using z_{orig} and $\hat{z}$ as defined in Section 3.1. The objective consists of four loss terms:

$$L_{\text{erase}} = \mathbb{E}_{\mathcal{D}_f}\left[\hat{z}^\top z_{\text{orig}} + \tfrac{1}{2}\hat{z}^\top \bar{t}_{\text{erase}}\right], \tag{2}$$

$$L_{\text{retain}} = \mathbb{E}_{\mathcal{D}_r}\left[1 - \hat{z}^\top z_{\text{orig}} + \tfrac{\beta}{d}\|\hat{z} - z_{\text{orig}}\|_2^2\right], \tag{3}$$

$$L_{\text{ctrl}} = \mathbb{E}_{\mathcal{D}_f}\left[1 - \hat{z}^\top \bar{t}_{\text{ctrl}}\right], \tag{4}$$

$$L_{\text{struct}} = \mathbb{E}_{\mathcal{D}_f \cup \mathcal{D}_r}\left[w(x) D_{\text{KL}}(p_{\text{orig}} \| p_{\text{adapted}})\right], \tag{5}$$

where $p_z = \text{softmax}(T_{\text{struct}} z/\tau_s)$ and $w(x) = 2$ for forget images and 1 for retain images. T_{struct} comprises peer and selected context anchors. L_{erase} encourages departure from the original representation and reduces erase-anchor similarity. L_{ctrl} guides the edited representation toward its declared ancestor, while L_{struct} preserves the relative structure of nearby anchors. Training introduces erase and retain losses first, then semantic control and structure preservation, followed by an auxiliary directional penalty (Appendix B.1). Its removal has negligible measured effect; it is not a claimed source of CSOA's empirical gains. To limit conflicting updates, each parameter tensor's forget gradient is projected away from a conflicting retain gradient before weighted combination with the other objectives (Yu et al., 2020). The ten-epoch schedule implements stages of 2/4/4 epochs; the full weights, projection equation, and optimizer are in Appendix B.5.

Retention-focused adaptation. CSOA-R augments the weighted retain objective by $\lambda_{\text{KD}} \mathbb{E}_{\mathcal{D}_r} D_{\text{KL}}(p^{\text{cls}}_{\text{orig}} \| p^{\text{cls}}_{\text{adapted}})$, where p^{cls} uses all canonical class texts at temperature 0.07. A fraction q of each retain batch is sampled from same-coarse peer classes within the existing retain training pool. The remaining CSOA objectives and gradient projection are retained. Validation selects $\lambda_{\text{KD}} = 4$, $q = 0.25$, and control weight 1.5, with the other weights at their original values. This variant targets retained-class utility. We evaluate its three changes in a complete factorial design and test its frozen transfer separately from fixed CSOA.

4 Experiments

4.1 Evaluation Metrics

Standard Forget Accuracy measures whether the target concept's top-1 prediction is suppressed in-domain. We show empirically that low forget accuracy can coexist with residual recognition or local utility damage, motivating four complementary measurements.

Bias Generalisation Gap (BGG). Let A_d and A'_d be original and edited target accuracy in domain d, and $e_d = (A_d - A'_d)/A_d$ when $A_d > 0$. We report $100(e_{\text{source}} - \text{mean}_d\, e_d)$ using a shared class gallery, alongside raw accuracies. Normalization accounts for the original model's domain capability; a zero gap alone is insufficient when neither domain is forgotten.

Synonym Erasure Rate (SER). The original score diagnostic is the minimum mean cosine decrease across specified aliases. To test whether alternative expressions actually recover the concept, we

additionally measure classification with five training-excluded descriptions under three templates, reporting average, worst-query, and originally-correct conditional residuals. The main comparison uses this direct recognition test; similarity changes are reported separately.

Hierarchy Control Score (HCS). HCS is the fraction of target images whose post-unlearning top-1 prediction selects the declared ancestor in the class gallery augmented by its normalized control embedding. It measures a specified semantic destination.

Peer Concept Stability (PCS). PCS is the original-to-edited KL divergence over peer-anchor distributions. We additionally report exact-label accuracy loss on the four same-coarse CIFAR classes, preventing a stable similarity distribution from masking lost neighboring-class recognition. These metrics evaluate complementary axes; their detailed definitions are in Appendix A.2.

4.2 Experimental Setup

Three evaluation scenarios stress-test the failure modes, with complete datasets and settings in Appendix B.3.

Visual bias and association leakage (Scene A). A single-source forget set from CIFAR-100 trains each method. Evaluation proceeds on its official test split and ImageNet (Deng et al., 2009), covering ten concepts in-domain and six with predefined cross-domain mappings. Each request uses 400 target and 4,096 retain training images; disjoint validation pools contain 100 and 1,000 images. Five seeds, a fixed final checkpoint, and unit adapter strength are used. The global zero-shot retain test covers the full non-target ImageNet validation pool. CSOA-R uses the same images and five seeds; one configuration is selected across all concepts using source validation data, as detailed in Appendix D.

Semantic collapse (Scene B). The original Oxford Pets hierarchy profiles (Parkhi et al., 2012) and Flowers-102 records (Nilsback & Zisserman, 2008) illustrate semantic redirection at multiple levels. We strengthen this scenario with controlled interventions on the ten Scene A concepts, independently measuring declared-ancestor preference and exact-label peer utility. Alternative roots and peer centroids test whether the original ancestor is the only useful endpoint.

Visual bias, cross-style (Scene C). Frozen CIFAR-trained or text-edited checkpoints are evaluated directly on DomainNet (Peng et al., 2019), without domain adaptation or target-style model selection. Six concepts and five seeds are tested on real, sketch, and clipart images. All official test images in a predeclared 26-class pool are retained; classification uses all 345 DomainNet class names and, separately, the original CIFAR gallery.

Implementation and hardware. CSOA trains 20,480 parameters at rank 16 on frozen CLIP ViT-B/32 (Radford et al., 2021). CLIP tasks run on single RTX 3090 or A6000 GPUs. Comparisons include GA, NegGrad+, SSD, SalUn, CLIPErase, SLUG, Text-image Space, and deterministic CCUP. A knowledge-matched control receives identical anchors, rank, splits, initialization, and training-step budget. Algorithm-specific settings and trainable components are disclosed in Appendix B.4.

4.3 Experimental Setup and Results

Table 1 reports source-domain forgetting, cross-domain forgetting, retained utility, and language-query robustness. Three key findings emerge. Text-image Space retains utility but leaves 27.07% ImageNet target residual. SLUG's selected zero edit is reported transparently and is not treated as an equally strong source-domain erasure. **(1) Low source forget accuracy does not settle robust concept erasure:** GA reaches 0% CIFAR target accuracy but retains 46.32% on ImageNet; CCUP reaches 0% in-domain while retaining 28.38% description recognition. **(2) Strong suppression can transfer its cost to retained utility:** CLIPErase achieves 0.00%/0.20% source/transfer target accuracy and 0.13% description residual, but loses 32.39 ImageNet retain points. **(3) CSOA provides a selective operating point:** target accuracies are 0.04%/0.07%, description residual is 1.10%, and source/transfer retain losses are 1.10/3.34 points. The knowledge-matched control also suppresses targets, but its source retain loss is 2.23 points, showing that semantic-resource access alone does not explain CSOA's selectivity. CSOA-R further reduces source/transfer retain losses to 0.08/1.40 points and peer loss to 0.37 points. Its description residual rises to 2.60%, but remains well below CCUP's 28.38%; this is a retention-oriented trade-off rather than uniform improvement over the fixed CSOA recipe. The fixed CSOA point favors stricter alternative-expression suppression and ancestor

Table 1: Scene A: CIFAR-100 (ten concepts) and ImageNet (six mapped concepts). Values are percentages, except peer accuracy loss in percentage points. Stochastic methods use five seeds; CCUP is deterministic. Description accuracy measures residual recognition under training-excluded expressions. CSOA-R is the retention-focused variant. SLUG selects a zero edit under the source retain constraint. Bold marks the best non-reference value, including displayed ties; n/a denotes an unavailable peer measurement.

Method	CIFAR-100 Fgt↓	CIFAR-100 Ret↑	ImageNet Fgt↓	ImageNet Ret↑	Descriptions ↓	Peer loss ↓
Original CLIP	73.10	61.49	78.97	59.70	34.16	0.00
GA	**0.00**	54.12	46.32	58.80	0.49	25.79
NegGrad+	32.10	60.77	76.01	59.59	15.16	3.47
SSD	69.04	60.07	78.37	57.89	27.67	1.31
SalUn	0.02	51.88	24.01	54.16	**0.04**	35.97
CLIPErase	**0.00**	**65.02**	0.20	27.31	0.13	9.13
CCUP	**0.00**	61.21	1.00	58.16	28.38	3.43
Text-image Space	3.22	61.52	27.07	59.53	4.30	n/a
SLUG	73.10	61.49	78.97	**59.70**	34.16	n/a
Same-knowledge control	**0.00**	59.27	**0.07**	55.40	0.22	14.61
CSOA	0.04	60.39	**0.07**	56.36	1.10	8.92
CSOA-R	0.22	61.41	0.93	58.30	2.60	**0.37**

Table 2: Fixed-recipe breadth beyond the primary ViT-B/32 comparison. ViT-B/16 and ViT-L/14 use ten concepts and five seeds; ViT-B/32 (40) uses 40 further concepts and two seeds with automatic anchors. Values are percentages except retain and peer losses (points). Descriptions are evaluated only for the ten original concepts. Bold compares CSOA and CSOA-R within each backbone.

Backbone	Method	Fgt↓	Ret loss↓	Peer loss↓	HCS↑	Desc↓
ViT-B/16	CSOA	**0.06**	0.58	5.99	**99.40**	**0.50**
	CSOA-R	0.12	**0.21**	**0.99**	97.66	0.55
ViT-L/14	CSOA	**0.00**	0.36	5.70	**99.04**	**0.05**
	CSOA-R	**0.00**	**0.02**	**0.62**	98.06	0.11
ViT-B/32 (40)	CSOA	0.24	1.48	3.69	**99.15**	n/a
	CSOA-R	**0.15**	**0.24**	**-0.41**	96.51	n/a

preference; CSOA-R is preferable when preserving neighboring and retained classes is the priority, provided its measured residuals meet the application's threshold.

Backbone and concept breadth. Table 2 shows that fixed CSOA and CSOA-R recipes suppress ten target concepts on ViT-B/16 and ViT-L/14. CSOA-R reduces neighboring-class loss while retaining high ancestor preference. On 40 further CIFAR-100 concepts with outcome-independent automatic anchors, CSOA-R has 0.15% mean target residual and 0.24-point retain loss, compared with 0.24% and 1.48 points for CSOA. This separate breadth test is not a matched baseline ranking. Frozen DomainNet tests on the six mapped concepts remain effective on both added backbones (Table 5).

Membership inference and score interpretation. Class-matched, calibration-selected attacks give near-chance membership AUC for fixed CSOA and CSOA-R. With 25 forget images, however, their post-edit attack AUCs rise to 0.654 and 0.691: behavioral suppression does not certify privacy removal (Appendix A.4). Ancestor redirection can also increase target-name cosine similarity while changing its class rank; recognition and absolute scores are therefore reported separately.

4.4 Hierarchical Control (Scene B)

The original Pets profiles (Appendix Figure 5) illustrate hierarchical redirection. For samoyed, target cosine similarity falls from 0.3384 to 0.2196 while mammal similarity rises from 0.2297 to 0.6229. Table 3 strengthens this evidence with matched interventions: removing only L_{ctrl} reduces HCS from 98.72% to 1.28% and increases description residual from 1.10% to 13.35%. Redirecting toward a

Table 3: Hierarchical-control interventions on the ten Scene A concepts, five seeds. HCS is always evaluated against the original declared ancestor, including alternative-endpoint rows. Peer loss is exact-label accuracy loss; PCS is forward KL.

Method	Fgt(%)↓	HCS(%)↑	PCS KL↓	Peer loss↓
CSOA	0.04	**98.72**	0.02	8.92
w/o L_{ctrl}	4.76	1.28	**0.01**	4.92
Broader root	**0.00**	78.82	0.02	15.86
Peer centroid	**0.00**	65.62	0.03	**2.21**

Table 4: Frozen DomainNet comparison: six concepts and the same 26-class image pool. Fgt is native-gallery target accuracy (%); loss is retain accuracy drop (pp). Five source seeds are used, except deterministic CCUP. Bold marks the lowest edited-model value per column. No target-style adaptation or model selection is performed.

	Real		Sketch		Clipart	
Method	Fgt↓	Loss↓	Fgt↓	Loss↓	Fgt↓	Loss↓
Original CLIP	94.39	0.00	86.52	0.00	85.49	0.00
CLIPErase	**0.00**	8.98	**0.00**	6.53	**0.00**	10.15
CCUP	**0.00**	1.08	**0.00**	1.39	**0.00**	1.11
SLUG	94.39	0.00	86.52	0.00	85.49	0.00
Text-image Space	8.37	**-0.12**	2.90	**-0.27**	14.98	**-0.46**
CSOA	**0.00**	2.72	**0.00**	2.19	**0.00**	3.32
CSOA-R	**0.00**	0.85	0.04	0.70	0.37	0.87

broader root or a peer centroid changes which semantics are preserved. The peer centroid reduces peer loss to 2.21 points but reaches only 65.62% preference for the original ancestor. Thus a declared ancestor is a controllable task choice; endpoint preference and preservation of neighboring classes are related but distinct objectives.

Generative evaluation diagnostics. Qwen2.5-VL-7B (Bai et al., 2025) exposes a measurement gap: ungated Pets Yes responses fall to 3.02%, while breed disclosure remains 91.83%. In a paired Qwen comparison on eight concepts and three styles, CSOA-R-VLM has lower direct-answer target disclosure and higher matched-class correctness than GA+KL and NPO+KL in every style. On real/sketch/clipart its disclosure is 5.00/5.42/7.57% and matched correctness is 84.29/75.83/75.06%. This advantage is prompt-dependent: open-description disclosure on clipart remains 23.54%, versus 11.32% for GA+KL. The same-checkpoint gate intervention also trades target coverage for retained utility (Appendix Table 28). On LLaVA-1.5-7B (Liu et al., 2024), a separate FP32 comparison finds 2.50% direct-answer residual for CSOA-R-VLM versus 52.50% for the adapted MMUnlearner objective, at 55.00% matched-class correctness for both. Parent-answer supervision in another paired LLaVA experiment suppresses measured target strings across three prompts on new disjoint images (Appendix E). The generative adaptations, budgets, and lexical scoring differ from the main CLIP protocol; these tests do not establish semantic deletion or uniform superiority.

4.5 Cross-Style Generalisation (Scene C)

Table 4 compares frozen edits under identical test images and galleries. CSOA suppresses native-gallery recognition in all three styles. CSOA-R reduces retain loss from 2.72/2.19/3.32 to 0.85/0.70/0.87 points on real/sketch/clipart, with small sketch/clipart residuals. Its retain cost is also lower than CCUP's 1.08/1.39/1.11 points, whereas CCUP achieves zero target residual. CLIPErase suppresses targets but incurs larger retain losses; Text-image Space preserves utility with substantially greater residual recognition. The source-selected SLUG checkpoint is unchanged, so its domain results do not constitute a comparison at matched source forgetting. Table 5 extends the frozen evaluation to two further backbones. CSOA-R reduces retain loss on all three ViT-B/16 styles while keeping target residual at or below 0.38%; on ViT-L/14 it gives zero target residual in each style and lower retain loss on real and sketch, with a 0.10-point higher loss on clipart.

Table 5: Frozen DomainNet evaluation on ViT-B/16 and ViT-L/14. Six mapped concepts, five seeds, the same 26-class test pool, and no target-style adaptation. Fgt is native-gallery target accuracy (%); loss is retain accuracy drop (pp). Bold compares CSOA and CSOA-R within each backbone.

		Real		Sketch		Clipart	
Backbone	Method	Fgt↓	Loss↓	Fgt↓	Loss↓	Fgt↓	Loss↓
ViT-B/16	CSOA	**0.00**	2.72	**0.00**	3.13	**0.00**	3.29
	CSOA-R	**0.00**	**0.94**	0.27	**1.71**	0.38	**1.50**
ViT-L/14	CSOA	**0.00**	0.96	**0.00**	0.83	**0.00**	**0.40**
	CSOA-R	**0.00**	**0.71**	**0.00**	**0.60**	**0.00**	0.50

Table 6: Component interventions at fixed final strength. The upper block uses ten concepts and five seeds; the lower block compares the same ten concepts and seeds 0–2, including its matched default. All retain/peer losses are percentage points.

Variant	CIFAR Fgt	Ret loss	ImageNet Fgt	Descriptions	HCS	Peer loss
CSOA (5 seeds)	0.04	1.10	0.07	1.10	98.72	8.92
Canonical erase	0.02	1.01	0.13	1.38	98.74	7.66
w/o L_{ctrl}	4.76	**0.07**	26.01	13.35	1.28	**4.92**
w/o L_{struct}	**0.00**	1.21	**0.00**	**0.31**	**98.76**	10.69
CSOA (matched 3 seeds)	**0.00**	**1.06**	**0.00**	1.06	98.67	**8.65**
w/o gradient projection	**0.00**	6.23	**0.00**	2.09	**100.00**	30.41
w/o curriculum	**0.00**	1.25	**0.00**	**1.02**	98.53	8.97

Holding the CIFAR gallery fixed gives CSOA-R residuals of 0.07/0.13/1.71%, compared with CSOA's 0/0/0.37% and CCUP's 0/0/0% (Appendix Table 18). For fixed CSOA, the normalized BGG on shared-gallery clipart is 0.48 points, with the original model's accuracy reported alongside it (Appendix C.5). Thus CSOA-R offers a lower-retain-cost transfer operating point, not uniformly stronger suppression. The strong original recognition and both galleries rule out interpreting all low residuals as inherited domain failure or a gallery effect.

Limited forget data. Appendix Figure 7 varies nested source subsets from 25 to 400 images without changing the recipe or optimization budget. With 25 images, CSOA-R has 0.60% source residual and 0.85-point peer loss, compared with 0% and 7.80 points for CSOA. CSOA-R's native-gallery DomainNet residuals remain 1.03/1.06/1.61%; increasing data reduces residuals while preserving its utility advantage. The worst source unit at 25 images retains 9% accuracy, so low macro averages do not imply uniform per-concept removal (Appendix D.3).

4.6 Ablation: One Component per Challenge

Table 6 pairs each component with the failure or utility cost it is intended to address. **QAS and association leakage:** replacing the erase bank by its canonical entry leaves source forgetting almost unchanged, but increases description residual from 1.10% to 1.38%. **Hierarchical control and semantic collapse:** removing L_{ctrl} collapses HCS and raises transfer residual to 26.01%. **Structure and neighboring utility:** removing L_{struct} increases peer loss from 8.92 to 10.69 points. **Gradient projection and selectivity:** on the matched three-seed set, its removal preserves zero source/transfer target accuracy but raises retain loss from 1.06 to 6.23 points and peer loss from 8.65 to 30.41 points. Enabling all losses from the first epoch has a smaller cost, raising ImageNet retain loss from 3.24 to 4.00 points.

Sensitivity. Varying loss coefficients by factors of 0.5 and 2 and ranks over 8/16/32 keeps source/transfer residual at or below 0.23%, with description residual of 0.88–2.47%. Lower control reduces retain cost but weakens redirection (Appendix C.2). The optional router's utility–coverage trade-off is documented separately in Appendix C.6.

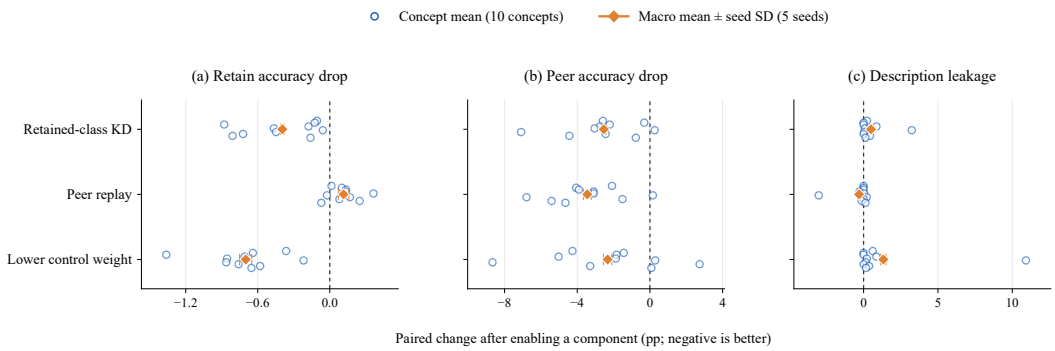

Figure 2: CSOA-R component effects in a full factorial study. Each component is enabled versus disabled under all four settings of the other two. Circles are concept means; diamonds are macro means with SD across five seed-level means. Negative is better.

Isolating CSOA-R's retention mechanisms. The full factorial study separates retained-class distillation, peer replay, and lower control weight (Figure 2; Appendix Table 22). Relative to full CSOA-R, removing distillation increases peer loss from 0.37 to 2.09 points; removing peer replay increases it to 1.88 points. Restoring control weight three raises source retain loss from 0.08 to 0.70 points, while increasing HCS from 94.98% to 98.60%. Lower control alone leaves 5.03-point peer loss, so it does not explain the combined variant's local preservation. The components have distinct trade-offs: peer replay favors neighboring classes, while stronger semantic control favors ancestor preference at greater retain cost.

Comparison scope. The CLIP baselines share test images, galleries, and a fixed source-only selection rule, but their trainable modules and optimization budgets differ. The tables compare disclosed operating points rather than a matched-budget Pareto frontier (Appendix B.4).

5 Discussion

Structural coupling of the three failure modes. Hierarchical control specifies the semantic endpoint, while structure preservation and gradient projection limit collateral changes. Knowledge-matched controls and factorial interventions distinguish resource access, semantic destination, and local preservation. The contribution concerns their joint control and evaluation; the auxiliary direction is not established as a source of distribution invariance.

Limitations. Finite semantic banks may under-cover open-world concepts, despite the automatic coarse-class protocol for 40 additional categories. The primary matched CLIP comparison covers ten concepts; the extra backbones share these concepts, while the 40-class extension uses two seeds and a different anchor policy. Each request edits one concept, so interference under simultaneous or sequential edits remains untested. Generative comparisons use distinct interventions and budgets, and lexical scores cannot establish correct semantic abstraction. Ancestor preference does not certify information removal or privacy protection.

6 Conclusion

We identified visual bias, association leakage, and semantic collapse in VLM concept unlearning. CSOA combines a multi-anchor specification with hierarchical control and conflict-aware retention. Matched interventions, three-backbone tests, a 40-concept extension, and paired generative comparisons support selective adaptation and CSOA-R's neighboring-class preservation, while exposing residual-recognition and privacy limits. Jointly evaluating suppression, semantic destination, and retained utility makes these trade-offs visible.

REPRODUCIBILITY STATEMENT

The original CSOA source supplies the model, anchors, and training losses. Appendix B.5 specifies the fixed recipe, Appendix B.4 the comparisons, and Appendices C.2 and D the sensitivity and selection protocols. The accompanying numeric snapshots and figure scripts reproduce the reported aggregates from saved records, for the completed experiments reported here. Model/checkpoint and prediction hashes connect the additional measurements to their fixed trained units.

ETHICS STATEMENT

The experiments use object-recognition datasets and evaluate behavioral concept suppression. They do not establish person-specific deletion or certified privacy removal. Permitted semantic ancestors are explicit task choices, and misleading substitutes must be distinguished from appropriate abstraction.

AI USE STATEMENT

AI Assistance: Yes, to aid or polish writing. Details are described in the paper. Yes, for generating synthetic datasets. Details are described in the paper. AI assisted manuscript wording, figure drafting, and the preparation of synthetic textual resources, including semantic descriptions and probes. Empirical plots were generated from saved experimental records. The textual resources and their use are documented in Appendix B.2 and the complete description bank in Appendix B.6. The authors are responsible for the accuracy of the manuscript and the reported experimental results.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Zikui Cai, Yaoteng Tan, and M. Salman Asif. Targeted unlearning with single layer unlearning gradient. In *International Conference on Machine Learning*, 2025. URL https://csiplab.github.io/slug/.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 463–480, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.07163.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12043–12051, 2024. doi: 10.1609/aaai.v38i11.29092.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Arpit Garg, Hemanth Saratchandran, and Simon Lucey. SineProject: Machine unlearning for stable vision-language alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Garg_SineProject_Machine_Unlearning_for_Stable_Vision-Language_Alignment_CVPR_2026_paper.pdf.

Jiahui Geng and Qing Li. SAUCE: Selective concept unlearning in vision-language models with sparse autoencoders. In *IEEE/CVF International Conference on Computer Vision*, pp. 3023–3033, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Geng_SAUCE_Selective_Concept_Unlearning_in_Vision-Language_Models_with_Sparse_Autoencoders_ICCV_2025_paper.pdf.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. MMUnlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7190–7206, 2025. doi: 10.18653/v1/2025.findings-acl.375. URL https://aclanthology.org/2025.findings-acl.375/.

Yejin Kim, Dongjun Hwang, Sungmin Cha, and Junsuk Choe. Knowledge vector weakening: Efficient training-free unlearning for large vision-language models. *arXiv preprint arXiv:2601.21794*, 2026. URL https://arxiv.org/abs/2601.21794.

Alexey Kravets and Vinay P. Namboodiri. Zero-shot class unlearning in CLIP with synthetic samples. In *IEEE/CVF 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*, 2025b. URL https://openreview.net/forum?id=V2SD2uVKEE.

Nilakshan Kunananthaseelan, Jing Wu, Trung Le, Gholamreza Haffari, and Mehrtash Harandi. DIET: Machine unlearning on a data-diet. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22698–22706, 2026. doi: 10.1609/aaai.v40i27.39431.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems*, pp. 1957–1987, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with MLLMU-Bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4105–4135, 2025. doi: 10.18653/v1/2025.naacl-long.207.

Ashish Mishra, Tarun Kumar, Gyanaranjan Nayak, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. Erasing clip memories: Non-destructive, data-free zero-shot class unlearning in clip models. *arXiv preprint arXiv:2512.14137*, 2025. URL https://arxiv.org/abs/2512.14137.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. URL https://www.robots.ox.ac.uk/~vgg/data/pets/.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019. URL https://ai.bu.edu/M3SDA/.

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

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*, pp. 30438–30452, 2025. doi: 10.18653/v1/2025.acl-long.1469. URL https://aclanthology.org/2025.acl-long.1469/.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

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

Jianing Zhu, Bo Han, Jiangchao Yao, Jianliang Xu, Gang Niu, and Masashi Sugiyama. Decoupling the class label and the target concept in machine unlearning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3327377ee37b8c87c43a7d6a4ea3d279-Abstract-Conference.html.

A Problem formulation and evaluation metrics

A.1 Formal Problem Formulation

Let $\mathcal{D}_f$ and $\mathcal{D}_r$ be disjoint forget and retain training sets, and let $\mathcal{T}_{c,d}$ contain test images of concept c in domain d. For a finite text gallery $\mathcal{G}$, prediction is $h_{\mathcal{G}}(x) = \arg\max_{t \in \mathcal{G}} \hat{z}(x)^\top t$. The task seeks low target accuracy on $\mathcal{T}_{c,d}$ for source and unseen domains, including alternative linguistic expressions of c; high preference for the declared allowed ancestor on target images; and preservation of non-target predictions. These requirements need not attain their optima simultaneously. An ancestor retains shared semantic attributes by design, whereas target recognition in a specified gallery measures whether the fine class remains identifiable. Accordingly, the experiments evaluate behavioral concept suppression and selective redirection, rather than a certificate that all information about c has been removed from every model component. Each request edits one concept; independent concept/seed runs are macro-averaged.

A.2 Metric Definitions

Forget and retain accuracy. Forget accuracy is the percentage of target test images whose predicted class remains the target. Retain accuracy is exact-label accuracy on non-target test images, and retain loss is original minus edited accuracy, in percentage points. The original model and edited model use the same images, preprocessing, and gallery within each paired comparison. A negative retain loss denotes improvement.

Bias Generalisation Gap. For source s and target domain d, let A_d and A'_d be original and edited target accuracies and $e_d = 1 - A'_d/A_d$. The normalized gap is $\mathrm{BGG}_d = 100(e_s - e_d)$; an average over target domains can also be reported. The denominator must be positive and the gallery must be shared across domains. We report raw original and edited accuracies alongside BGG, since a small gap can otherwise describe two equally ineffective edits. When $A'_s = 0$, BGG reduces to $100A'_d/A_d$. This ratio of accuracies differs from the conditional residual on the subset of images originally classified correctly.

SER and recognition under alternative expressions. For specified alias vectors t_i, the original similarity diagnostic is $\mathrm{SER} = \min_i \mathbb{E}_{x \in \mathcal{T}_c}[z_{\mathrm{orig}}(x)^\top t_i - \hat{z}(x)^\top t_i]$. Despite its historical name, this is a cosine decrease, not a percentage of forgotten images. The main-table description metric instead replaces the target class entry by one external description and measures edited exact-label recognition. For each run we average the 15 query accuracies (five descriptions times three templates); worst-query accuracy is their maximum. For each query, the originally-correct conditional residual is the fraction of originally correct target images that remain correct after editing; queries with no originally correct image have no conditional denominator. All-query residuals retain all queries.

HCS and PCS. The HCS gallery contains the original 100 CIFAR class vectors and one normalized declared control vector. HCS is the percentage of target images selecting that control entry. For root/peer-centroid interventions, evaluation still uses the original declared ancestor, allowing a direct comparison of destinations. PCS is $\mathbb{E}_x D_{\mathrm{KL}}(p_{\mathrm{orig}}(x) \| p_{\mathrm{edited}}(x))$ over four same-coarse peer text vectors at temperature 0.07. Exact-label peer loss is independently measured on the 400 test images belonging to those four classes, using the standard 100-class gallery. The latter can reveal damage even when the peer-distribution KL is small.

A.3 Evaluation Metric Design and Validity

The original framework distinguishes three failure modes, but a score on one axis cannot certify success on the others. The description bank tests finite held-out expressions rather than universal language robustness. HCS tests the declared semantic destination, while peer accuracy tests non-target examples. Neither metric implies privacy deletion or full generative semantic correctness. The CIFAR/ImageNet comparison holds the model and its edit fixed across tests; DomainNet additionally exposes sensitivity to gallery composition. Main tables show every method under its disclosed fixed recipe, including retain costs and residuals that prevent a uniform dominance claim. Historical hierarchy profiles remain separate from the matched multi-seed experiment.

A.4 PRIVACY SCOPE

We audit membership in the adapter's forget training set, not membership in CLIP's unknown pretraining data. The primary non-member population consists of official CIFAR test images of the same class. A secondary population uses excluded source-training images, which also served earlier validation and selection; this is a sensitivity analysis rather than an untouched non-member benchmark. Exact saved indices identify both populations. At most 100 members and 100 non-members are sampled per unit; the low-data conditions use all available members.

Each population is split into disjoint calibration and evaluation halves. Calibration chooses the attack signal and its direction; evaluation reports AUC with that direction held fixed, without flipping sub-chance values after observing test labels. The post-edit threat model considers target log probability, negative entropy, correctness, and logit margin. The paired-access threat model additionally considers the change in target log probability and feature cosine shift. Log probabilities are computed stably in float64. Original and edited models are audited separately under the same population splits.

Table 7 includes fixed CSOA/CSOA-R, six validation-selected operating points, and the three reduced-data budgets. The full audit covers 580 checkpoint units. Means first average concepts within seed and then five seeds; fixed/selected recipes cover ten concepts, while reduced-data recipes cover the six mapped concepts. Consequently, changes between these blocks are not purely sample-size effects. The near-chance original-model attacks provide no evidence of a successful privacy-removal intervention. Conversely, the elevated AUC at 25 images shows that the adaptation itself can leave a detectable training-set signature despite strong functional forgetting. These measurements do not establish deletion relative to a retraining oracle, nor do they rule out stronger attacks. Their scope is adaptation-set membership under the disclosed finite attacks. Source-test forgetting, ancestor preference, and privacy are distinct claims.

Table 7: Forget-training membership AUC against class-matched official-test non-members. Values are mean $\pm$ SD across five seed means. Attack selection and direction use only calibration data. Chance is 0.5; this diagnostic table does not assign a performance winner. Parentheses indicate forget training images.

	Post-edit signals		Paired-access signals	
Operating point	Original	Edited	Original	Edited
CSOA (400)	0.492 ± 0.021	0.490 ± 0.015	0.486 ± 0.028	0.490 ± 0.015
CSOA-R (400)	0.492 ± 0.021	0.491 ± 0.006	0.486 ± 0.028	0.488 ± 0.007
Loss-weight variant	0.492 ± 0.021	0.505 ± 0.012	0.486 ± 0.028	0.506 ± 0.014
Rank/LR variant	0.492 ± 0.021	0.503 ± 0.011	0.486 ± 0.028	0.504 ± 0.011
Retention/peer variant	0.492 ± 0.021	0.491 ± 0.006	0.486 ± 0.028	0.488 ± 0.007
Control-margin variant	0.492 ± 0.021	0.492 ± 0.006	0.486 ± 0.028	0.498 ± 0.014
Erase-margin variant	0.492 ± 0.021	0.508 ± 0.017	0.486 ± 0.028	0.501 ± 0.024
Ridge initialization	0.492 ± 0.021	0.506 ± 0.013	0.486 ± 0.028	0.505 ± 0.017
CSOA (25)	0.498 ± 0.020	0.654 ± 0.025	0.491 ± 0.032	0.635 ± 0.035
CSOA (50)	0.508 ± 0.022	0.548 ± 0.043	0.515 ± 0.011	0.545 ± 0.045
CSOA (100)	0.501 ± 0.012	0.505 ± 0.011	0.507 ± 0.003	0.514 ± 0.018
CSOA-R (25)	0.498 ± 0.020	0.691 ± 0.027	0.491 ± 0.032	0.688 ± 0.029
CSOA-R (50)	0.508 ± 0.022	0.567 ± 0.028	0.515 ± 0.011	0.566 ± 0.023
CSOA-R (100)	0.501 ± 0.012	0.529 ± 0.017	0.507 ± 0.003	0.523 ± 0.020

B IMPLEMENTATION AND SEMANTIC RESOURCES

B.1 CONCEPT-DIRECTION CONSTRUCTION AND SEMANTIC RESOURCES

The evaluated original-source implementation constructs a direction from semantic anchors and source training features. It does not fit a multi-domain visual subspace using DomainNet test images. The framework illustration and empirical claims concern the source-only construction specified below. Its sequential projections reduce interference along the supplied directions, but do not guarantee simultaneous orthogonality to all of them because the candidate directions are not orthogonalized

Table 8: Membership sensitivity analysis using class-matched held-out source-training images as non-members. These images were available for earlier validation/selection. This panel is kept separate from Table 7; aggregation and attack selection are otherwise identical.

Operating point	Post-edit signals Original	Post-edit signals Edited	Paired-access signals Original	Paired-access signals Edited
CSOA (400)	0.496 ± 0.014	0.515 ± 0.014	0.500 ± 0.006	0.514 ± 0.009
CSOA-R (400)	0.496 ± 0.014	0.498 ± 0.015	0.500 ± 0.006	0.496 ± 0.019
Loss-weight variant	0.496 ± 0.014	0.512 ± 0.020	0.500 ± 0.006	0.509 ± 0.030
Rank/LR variant	0.496 ± 0.014	0.515 ± 0.022	0.500 ± 0.006	0.508 ± 0.029
Retention/peer variant	0.496 ± 0.014	0.498 ± 0.015	0.500 ± 0.006	0.496 ± 0.019
Control-margin variant	0.496 ± 0.014	0.517 ± 0.011	0.500 ± 0.006	0.504 ± 0.013
Erase-margin variant	0.496 ± 0.014	0.519 ± 0.016	0.500 ± 0.006	0.519 ± 0.013
Ridge initialization	0.496 ± 0.014	0.512 ± 0.013	0.500 ± 0.006	0.506 ± 0.011
CSOA (25)	0.491 ± 0.031	0.635 ± 0.047	0.492 ± 0.054	0.637 ± 0.041
CSOA (50)	0.507 ± 0.014	0.562 ± 0.039	0.514 ± 0.029	0.559 ± 0.036
CSOA (100)	0.497 ± 0.016	0.542 ± 0.023	0.497 ± 0.009	0.537 ± 0.032
CSOA-R (25)	0.491 ± 0.031	0.674 ± 0.038	0.492 ± 0.054	0.670 ± 0.038
CSOA-R (50)	0.507 ± 0.014	0.585 ± 0.029	0.514 ± 0.029	0.583 ± 0.027
CSOA-R (100)	0.497 ± 0.016	0.552 ± 0.023	0.497 ± 0.009	0.544 ± 0.030

jointly. The method uses this direction only as an auxiliary geometric preference. The frozen transfer measurements do not identify it as the cause of cross-domain suppression.

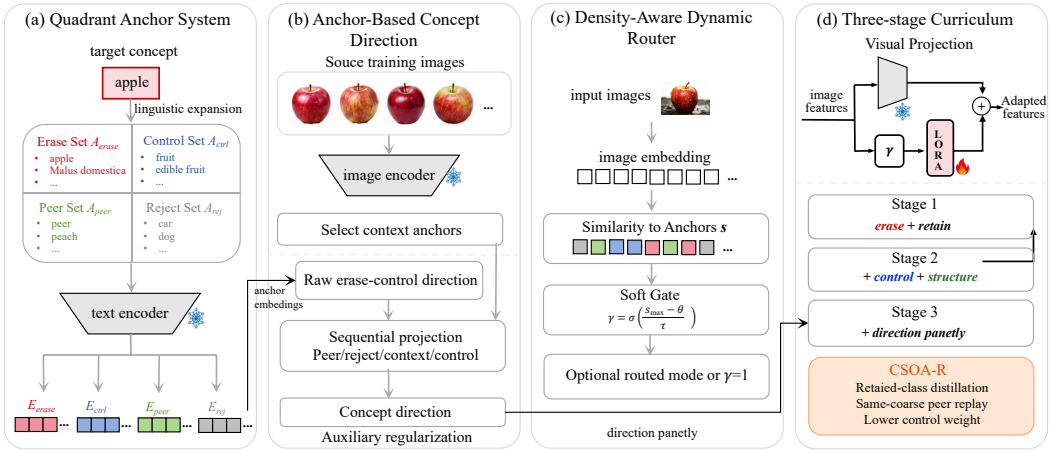

Figure 3: Complete implementation schematic. Panels (a), (b), and (d) describe the semantic banks, auxiliary direction, and training schedule. Panel (c) shows an optional density-aware router; the primary CLIP protocol sets $\gamma = 1$ and does not depend on routing. CSOA-R adds retained-class distillation, same-coarse peer replay, and lower control weight.

The final training stage also uses the directional regularizer

$$L_{\text{dir}} = \frac{1}{dr} \|W_B - d_c d_c^\top W_B\|_F^2. \tag{6}$$

where d_c is the normalized anchor-derived direction. This term expresses an auxiliary preference, rather than a proven cross-domain invariance mechanism.

B.2 Semantic Resources

Finite training banks. The implementation uses explicit 15-string erase banks for apple and rose and a fixed 15-template fallback for the other eight evaluated classes. For a name c, the fallback contains the canonical name and constructions such as "c object," "real c," "close-up of c," and "c on white background." It is not a complete synonym inventory. Apple and rose additionally have

explicit peer banks; fallback classes have no explicit training peer bank and their structure loss uses the selected context anchors. The four-peer diagnostic set is instead defined independently by official CIFAR coarse labels for *every* concept. Training peers and diagnostic peers must not be conflated. AI-assisted synthetic textual resources specify the semantic query space. They are distinct from benchmark image samples; the complete external-description bank is given below.

Anchor construction. Erase, initial control, peer, and rejection strings are encoded with the source's photo-prefix convention, except strings already beginning with "a" or "an." The declared Scene-A control bank is then encoded directly and its normalized mean replaces the initial control mean. Context candidates are encoded directly. The mean feature of the first 16 saved forget-training indices selects the five highest-similarity context candidates, and the direction and structure matrix are recomputed. The ordered interference directions are the rejection mean, context mean, the first two right singular vectors of the peer-text matrix when present, and the declared control mean. For normalized u_j, the recursion is $v_{j+1} = v_j - (v_j^\top u_j)u_j$, starting with $v_0 = \bar{t}_e - u_c$, followed by normalization. The u_j are not jointly orthogonalized. Consequently, the last projection can reintroduce a component along an earlier direction; exact joint orthogonality or a visual-domain invariance theorem is not assumed.

Table 9: Declared original control banks. The diagnostic ancestor is the normalized mean of the encoded phrases.

Target	Control phrases
Apple	fruit; produce; food
Bicycle	vehicle; transport
Butterfly	insect; arthropod; animal
Elephant	large mammal; mammal; animal
Lawn mower	machine; tool; equipment
Mountain	landform; terrain; landscape
Rose	flower; plant; organism
Tiger	big cat; wild cat; mammal; animal
Whale	marine mammal; mammal; animal
Woman	person; human

Description scoring. The three templates are "a photo of {description}," "an image of {description}," and "this image depicts {description}." For each template, the other 99 class names use the same template, and only the target class entry is replaced by the held-out description. This is an exact-label classification probe, not free-form semantic understanding or a standard image-retrieval benchmark. The diagnostic runner checks literal non-overlap with training erase strings. Conditional residuals exclude zero-denominator queries; all-query edited accuracy retains all 15 queries. Description-query means are computed within a concept/run, then averaged across concepts and seeds. The table source exposes both unconditional and conditional measures.

Semantic diagnostics. Ancestor preference uses an augmented 101-entry gallery. Peer accuracy uses the ordinary 100-entry class gallery and four same-coarse fine classes, with 100 test images per class. A model can redirect target images well while harming these non-target images; the two metrics evaluate different examples. Peer feature cosine compares original and adapted unit representations. The peer-distribution diagnostic uses the four peer text vectors at temperature 0.07 and measures $D_{\mathrm{KL}}(p_{\mathrm{original}} \| p_{\mathrm{edited}})$. These quantities are not membership-inference scores.

B.3 Experimental Setup

Data and requests. Each CLIP run edits one concept, rather than jointly removing all ten concepts. The 400/100 target training/validation split is drawn from the 500 official CIFAR training images of the class. The 4,096/1,000 retain training/validation pools are disjoint subsets of non-target training images. The official test set is separate from these pools. Concepts and seeds have equal weight in the aggregate; ImageNet is averaged only over apple, bicycle, butterfly, elephant, tiger, and whale. The remaining concepts have no mapping in the frozen evaluation resource and are excluded

from ImageNet aggregates. The ImageNet retain evaluation covers all non-mapped classes in the 50,000-image validation set, not just the source-domain retain pool.

B.4 Comparison implementations and knowledge-matched controls

The image-trained comparison methods use the same CIFAR train/validation index partitions and test galleries as CSOA. CCUP and Text-image Space adaptation use text rather than image training data. The comparison settings below are fixed across concepts and seeds. Matched data does not imply matched optimization budget or identical trainable parameter subsets. In particular, these comparisons should not be read as the optimum attainable by official implementations or as a faithful reproduction of every published training protocol.

Table 10: Fixed comparison settings. Image-trained methods receive 400 forget and 4,096 retain images; CCUP uses text embeddings only.

Method	Settings
GA	Learning rate 10^{-5}; 5 epochs; retain weight 0.
NegGrad+	Input learning rate 10^{-5}, internally scaled to 3×10^{-6}; 5 epochs; retain weight 8.
SSD	Dampening $\alpha = 0.05$; selection ratio 0.2; retain-ratio selection; all forget batches and 20 retain batches for estimation.
SalUn	Saliency ratio 0.10; learning rate 2×10^{-5}; 10 epochs; retain weight 100.
CLIPErase	Adam, learning rate 10^{-6}; 10 epochs; both image and text encoders; six-term published objective.
CCUP	Closed-form projection; $\lambda = 100$, $\mu = 1$; no image training.

Implementation differences. GA and retain-regularized NegGrad+ update the visual projection and post-layer normalization with AdamW (zero weight decay). The latter applies an internal factor of 0.3 to the supplied learning rate. The SSD adaptation thresholds a forget/retain importance ratio at a chosen fraction and multiplies selected weights by $1 - 0.05$. The SalUn adaptation uses a saliency mask, random-label training, and retained-feature MSE, with a cosine learning-rate schedule and weight decay 10^{-4}. These task adaptations are not identical to all details of the cited original algorithms.

CLIPErase and CCUP. CLIPErase uses the published six-term forgetting, retention, and cross-modal-consistency objective, updates both encoders, and uses Adam at 10^{-6} for ten epochs. Its training pairs are the same CIFAR images with canonical class-caption text; this is a class-unlearning task adaptation, not replication of every original pretraining or data protocol. We retain the fixed configuration and report all fifty runs. Image-level predictions, loss curves, and final-state hashes are saved; the thirty six-concept states needed for frozen DomainNet evaluation were subsequently reconstructed under the fixed recipe, matching their historical full-state hashes and source predictions exactly. CCUP (Mishra et al., 2025) uses the published closed form $P = (I + \mu T_r T_r^\top)(I + \lambda T_f T_f^\top + \mu T_r T_r^\top)^{-1}$, with fixed $\lambda = 100$, $\mu = 1$, normalized canonical class text vectors, and visual projection $W' = PW$. It uses no training images and has one deterministic result per concept. The same CIFAR/ImageNet galleries, exact-label peers, and external descriptions evaluate both methods. The supplementary numerical snapshot includes the ten per-concept CCUP records and fifty CLIPErase records; neither is selected on test outcomes.

SLUG and Text-image Space. SLUG (Cai et al., 2025) uses forget/retain gradients, Pareto layer selection, and a single gradient edit. Our Hugging Face port selects the deepest Pareto layer and chooses step magnitude using source validation under a three-point retain-loss cap. The available candidate set selects zero magnitude in all fifty runs. We report this outcome rather than omit failed forgetting, but it does not establish the best performance attainable by SLUG or a comparison at matched suppression. Text-image Space (Kravets & Namboodiri, 2025b) uses rank five, learning rate 0.01, and 2,000 steps per text-only trial, with initial forget/preserve/weight coefficients 1.1/0.3/1.0 and the published reduction threshold 0.7. It adapts the shared space; test text galleries are re-encoded using the edited text model. Both methods use ten concepts and five seeds. Their frozen domain

tests use the six mapped concepts and identical image indices; the reconstructed states match the source-state hashes, including the corrected representation of signed zero for SLUG's no-op states. Exact-label peer metrics were not available for these two source comparisons and are not imputed from global retain scores.

Same-knowledge margin control. This comparator uses the same A/B initialization, rank, semantic banks, train splits, paired-batch sequence, 320 optimization steps, and cosine learning-rate schedule as the corresponding full CSOA run. Its erase term is a canonical class margin plus half an alias-versus-control margin:

$$L_{\text{class}} = \mathbb{E}_{\mathcal{F}}\left[\max\left(z'^{\top}t_c - \max_{j\neq c} z'^{\top}t_j + 0.05, 0\right)\right],$$
$$L_{\text{alias}} = \mathbb{E}_{\mathcal{F}}\left[\max\left(\max_{t\in T_e} z'^{\top}t - z'^{\top}u_c + 0.05, 0\right)\right].$$

Its weighted objective uses erase $2(L_{\text{class}} + 0.5L_{\text{alias}})$, retain weight 8 with MSE coefficient 0.25, control weight 2, and unweighted-per-example structure KL with weight 1. Control and structure terms begin in epoch 4. It uses ordinary backward on the summed loss, no directional penalty, and a $3/4/3$ stage convention. Thus the test controls knowledge access and the training budget, but changes the optimization recipe as a package; it is not a pure PCGrad or single-coefficient ablation.

Single interventions. The no-control and no-structure variants set only the corresponding coefficient to zero in the original CSOA trainer. The joint no-hierarchy variant removes control, structure, and directional penalties together. Canonical erase retains the control and contextual resources and changes only the erase bank to a canonical entry. Alternative redirection replaces the declared control bank with broader roots or same-coarse peer class names. Downstream quantities depending on the control mean are recomputed, including the directional construction. The peer centroid is the normalized mean of four class-name text vectors, not a learned visual centroid or one arbitrarily chosen sibling label.

B.5 Implementation hyperparameters

Table 11: Fixed CSOA training configuration. CSOA-R changes the control weight and adds retention distillation and peer replay, as specified in Appendix D.

Backbone / trainable component	CLIP ViT-B/32 / visual projection LoRA
Feature dimensions / rank	$768 \to 512$ / 16
Trainable parameters	$16(768 + 512) = 20{,}480$
Initialization	$A_{ij} \sim \mathcal{N}(0, 0.01^2)$, $B = 0$
Optimizer	AdamW, learning rate 3×10^{-4}
Weight decay / gradient norm cap	10^{-4} / 2
Schedule	CosineAnnealingLR over ten epochs
Paired batches	32 per epoch; 64 forget + 64 retain images
Loss weights (e, r, c, s, o)	$(2, 6, 3, 1, 0.5)$
Retain MSE coefficient / structure temperature	0.3 / 0.07
Epochs with erase + retain	1–2
Epochs additionally with control + structure	3–6
Epochs additionally with directional penalty	7–10
Primary checkpoint and strength	Epoch 10; strength $s = 1$
Primary train/evaluation routing	Always on

Stage boundaries and gradients. The stage function evaluates progress as epoch$/10$, with epochs numbered 1–10, and uses strict comparisons with 0.3 and 0.7. This yields stages $(1, 1, 2, 2, 2, 2, 3, 3, 3, 3)$, rather than rounding the nominal proportions to $3/4/3$ epochs. For each parameter tensor, the code computes $g_e = \nabla L_e$ and $g_r = \nabla L_r$ and replaces

$$g_e \leftarrow g_e - \frac{\min(0, \langle g_e, g_r\rangle)}{\|g_r\|^2 + 10^{-12}} g_r.$$

The final gradient combines the weighted projected forget gradient, retain gradient, and active control/structure/directional gradients, followed by clipping and AdamW. This is a tensor-wise forget-versus-retain projection, not a randomized projection across all task pairs. The scalar control loss is $1 - \cos$; it does not stop once a margin is reached. The KL direction is original-to-edited. Both distinctions matter when reproducing the method.

Cached features and checks. The frozen visual backbone permits caching pre-projection features without changing the adapter's function. The runner compares cached and raw-image features on recorded samples and checks the public adapter forward against the cache-based implementation. Each training artifact records the original source hashes, split identity, initialization hash, final epoch, and weights. The main protocol uses fixed final strength one; additional validation-selected strengths are separate operating points. An audit confirmed matching splits and initialization for the 250 same-knowledge and isolated-control training units against their corresponding CSOA references. The additional exact-label diagnostics read frozen checkpoints and perform no training.

Scope of the empirical design. The fixed final-checkpoint rule avoids choosing a different epoch or strength for each reported test metric. The concept set, bank, and recipe were developed through earlier experiments, however, so the study is not a fresh confirmation on ten previously unseen development concepts. External descriptions are absent from the training erase bank; this is a statement about training-query exclusion, not a claim that no researcher had previously examined any related category.

B.6 Complete external-description bank

The following five descriptions per concept are fixed before evaluating the additional final sensitivity experiments. Each is used with the three templates in Appendix B.2. All descriptions are retained regardless of their test outcomes.

Apple (1) the crisp pome represented by Gala and Fuji cultivars; (2) a red-or-green orchard pome whose juice is fermented into cider; (3) the temperate fruit whose varieties include Honeycrisp and Pink Lady; (4) a round tree fruit with a central core typified by Granny Smith; (5) the firm-fleshed pome sold in Golden Delicious and Braeburn varieties.

Bicycle (1) a pedal-powered two-wheeled vehicle; (2) a human-powered cycle with handlebars and pedals; (3) a two-wheeler propelled by a chain and pedals; (4) a non-motorized vehicle ridden by pedaling; (5) a cycle frame mounted on two inline wheels.

Butterfly (1) a diurnal lepidopteran with clubbed antennae; (2) a day-active lepidopteran that rests with its wings upright; (3) a club-antennaed insect emerging from a chrysalis; (4) a colorful lepidopteran with broad patterned wings and clubbed antennae; (5) the day-flying winged stage after a caterpillar chrysalis.

Elephant (1) a massive gray proboscidean with a trunk and tusks; (2) the largest living land mammal; (3) a thick-skinned herbivore with a prehensile trunk; (4) a large-eared pachyderm with ivory tusks; (5) a herd-forming proboscidean that uses its trunk to grasp objects.

Tiger (1) a striped orange-and-black big cat; (2) a large Asian feline with dark vertical stripes; (3) the striped member of the genus Panthera; (4) a solitary striped feline apex predator; (5) a Panthera cat distinguished by black stripes and orange fur.

Lawn mower (1) a wheeled machine that cuts turf with a rotating blade; (2) a powered garden device used to trim grass evenly; (3) a push-operated landscaping machine with a grass-cutting deck; (4) a motorized yard tool that shortens turf; (5) the garden machine rolled across grass to keep it uniformly short.

Mountain (1) a high natural landform rising steeply above surrounding terrain; (2) a large rocky elevation with a summit and slopes; (3) an elevated landmass whose upper point forms a peak; (4) a

towering geological formation climbed toward its summit; (5) a steep highland feature extending far above nearby ground.

Rose (1) a fragrant thorn-stemmed flower with layered petals; (2) the ornamental bloom commonly given on romantic occasions; (3) a garden flower borne on a prickly woody stem; (4) a many-petaled fragrant blossom from a thorny shrub; (5) the cut flower often arranged in a romantic bouquet.

Whale (1) a giant fully aquatic cetacean that breathes through a blowhole; (2) a massive ocean-dwelling mammal with tail flukes; (3) a marine cetacean that surfaces to breathe air; (4) a huge warm-blooded sea animal propelled by horizontal flukes; (5) the ocean mammal whose species include blue and humpback.

Woman (1) an adult female human; (2) a female person who has reached adulthood; (3) an adult human described by the female gender; (4) a grown female person; (5) the adult female member of the human species.

C Additional CLIP results

C.1 Backbone and concept breadth

The final breadth test applies the same fixed CSOA and CSOA-R recipes to OpenAI CLIP ViT-B/16 and ViT-L/14 on the original ten concepts and five historical training/validation splits. No backbone-specific or test-style selection is used. The five completed lanes contain 360 concept–seed–method units in total. The two added backbones share the original ten concept names, while the separate ViT-B/32 extension predeclares two additional concepts from each of 20 CIFAR-100 coarse classes, excludes the original ten, and uses two seeds. Its 40-class protocol constructs controls from the coarse-class name and peers from the four other classes in the same coarse group; it does not claim that these automatic anchors have the same quality as the original manual banks.

The CIFAR metadata calls numeric class 72 *seal*, whereas the historical CSOA source class-text list calls that index *sea urchin*. This breadth test uses the metadata name *seal* for its 100-class gallery without changing the archived source code. All ten original target IDs and their saved splits pass identity checks, but absolute gallery-based scores should not be treated as bitwise replications of the earlier ViT-B/32 table. The 40 additional concepts use deterministic fresh source-training splits. Table 2 reports the complete concept macro averages; its 40-class worst concept mean target residual is 9.0% for CSOA and 1.5% for CSOA-R, so the averages do not imply uniform zero residual.

C.2 Sensitivity and training mechanisms

The sensitivity study uses ten fixed interventions, ten concepts, and seeds 0–2. Each intervention is compared to the corresponding saved original-CSOA run with matching concept and seed. The five job groups reproduce the same reference summaries, rather than contributing independent repetitions of the reference. The default settings are $\lambda_c = 3$, $\lambda_s = 1$, $\lambda_r = 6$, and rank 16. We retain these original values in the primary comparison and report all additional settings without selecting a winner from test results. The no-projection intervention disables only the forget/retain gradient projection; the no-curriculum intervention activates all losses from epoch one. Source-code identities, final epoch, prediction hashes, and aggregate consistency were checked for all units.

Table 12: All sensitivity and training-mechanism settings, mean ± sample SD over three seed-level concept averages. Fgt and descriptions are percentages; retain and peer losses are percentage points. ImageNet uses six mapped concepts.

Setting	CIFAR-100 Fgt	CIFAR-100 Ret loss	ImageNet Fgt	ImageNet Ret loss	Descriptions	Peer loss
Original CSOA	**0.00±0.00**	1.06±0.06	**0.00±0.00**	3.24±0.09	1.06±0.13	8.65±0.42
$\lambda_c = 1.5$	0.07±0.12	**0.24±0.02**	**0.00±0.00**	**1.07±0.04**	2.47±0.23	**5.08±0.28**
$\lambda_c = 6$	0.20±0.20	2.21±0.07	0.22±0.19	4.13±0.10	1.16±0.10	11.93±0.79
$\lambda_s = 0.5$	**0.00±0.00**	1.08±0.05	**0.00±0.00**	3.25±0.09	**0.88±0.14**	8.88±0.36
$\lambda_s = 2$	0.07±0.06	1.04±0.06	0.11±0.19	3.24±0.09	1.19±0.15	8.43±0.42
$\lambda_r = 3$	**0.00±0.00**	1.20±0.11	**0.00±0.00**	3.60±0.20	1.02±0.17	9.34±0.79
$\lambda_r = 12$	0.07±0.06	0.83±0.02	**0.00±0.00**	2.78±0.05	1.07±0.06	6.98±0.62
$r = 8$	**0.00±0.00**	1.03±0.07	**0.00±0.00**	2.77±0.07	1.46±0.19	9.81±0.36
$r = 32$	0.03±0.06	0.97±0.03	**0.00±0.00**	3.53±0.17	1.40±0.04	5.88±0.19
w/o gradient projection	**0.00±0.00**	6.23±0.08	**0.00±0.00**	8.60±0.14	2.09±0.02	30.41±0.57
w/o curriculum	**0.00±0.00**	1.25±0.04	**0.00±0.00**	4.00±0.05	1.02±0.05	8.97±0.28

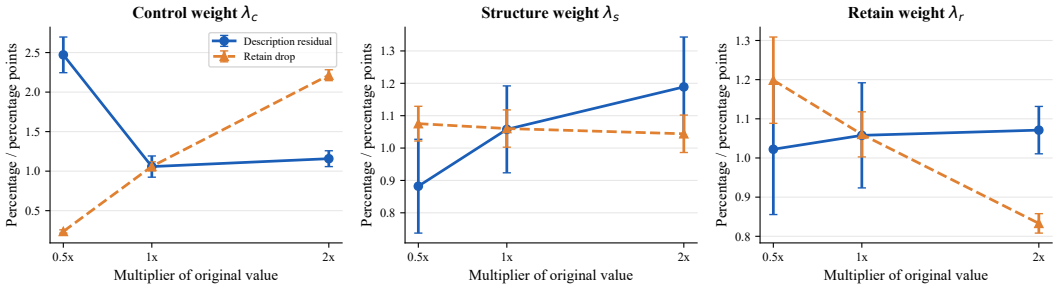

Figure 4: Sensitivity to control, structure, and retain weights. All other settings remain at the original values. Error bars are sample SD across the three seed-level macro averages.

The coefficient/rank interventions maintain low source and transfer target accuracy over the tested range. In contrast, removing gradient projection increases CIFAR retain loss by 5.17 points and peer loss by 21.76 points while maintaining zero target accuracy. This distinguishes the mechanism's contribution to selectivity from its contribution to target suppression. Activating all losses immediately has a smaller effect, increasing ImageNet retain loss by 0.76 points; it does not cause forgetting to fail.

C.3 Full aggregates and peer stability

For a metric $m_{c,s}$, we first average over eligible concepts for each seed, $\bar{m}_s = |\mathcal{C}|^{-1} \sum_c m_{c,s}$. The uncertainty tables report the mean and sample standard deviation of the five $\bar{m}_s$ values. This describes sensitivity to the five training seeds for the specified concept set; it is not an estimate of uncertainty over all possible concepts. ImageNet excludes unmapped concepts. A double hyphen denotes a quantity not included in the corresponding diagnostic aggregate, not zero error. CCUP is deterministic and is shown without a seed standard deviation.

The effect on neighboring classes varies by concept; global retain accuracy should not be substituted for these local utility measurements. Alternative control endpoints also change this balance: a peer centroid lowers peer loss but is less likely to select the original ancestor. This comparison supports an explicit semantic target chosen for the task, rather than the claim that every ancestor is uniformly optimal.

C.4 Pets and Flowers hierarchy profiles

The original Scene B archive contains twelve runs: two Pets breeds and two Flowers classes, each with three declared control levels. Figure 5 uses the original and CSOA cosine profiles from the

Table 13: Source and language results, mean ± sample SD over five seed-level concept averages. Percentages for target/description, percentage points for retain loss.

Method	CIFAR target	CIFAR retain drop	External description
GA	**0.00 ± 0.00**	7.38 ± 0.22	0.49 ± 0.03
NegGrad+	32.10 ± 0.14	0.72 ± 0.01	15.16 ± 0.05
SSD	69.04 ± 0.82	1.42 ± 0.06	27.67 ± 0.33
SalUn	0.02 ± 0.04	9.61 ± 0.66	0.04 ± 0.06
Same-knowledge margin	**0.00 ± 0.00**	2.23 ± 0.03	0.22 ± 0.01
CSOA	0.04 ± 0.05	1.10 ± 0.07	1.10 ± 0.11
Canonical erase	0.02 ± 0.04	1.01 ± 0.08	1.38 ± 0.17
No hierarchy (joint)	3.46 ± 1.18	0.10 ± 0.02	10.79 ± 1.02
No L_{ctrl}	4.76 ± 1.27	0.07 ± 0.03	13.35 ± 0.88
No L_{struct}	**0.00 ± 0.00**	1.21 ± 0.08	0.31 ± 0.06
Broader root	**0.00 ± 0.00**	1.56 ± 0.08	**0.00 ± 0.00**
Peer centroid	**0.00 ± 0.00**	1.00 ± 0.12	**0.00 ± 0.00**
CLIPErase	**0.00 ± 0.00**	**−3.53 ± 0.58**	0.13 ± 0.29
CCUP	**0.00**	0.29	28.38
CSOA-R	0.22 ± 0.04	0.08 ± 0.05	2.60 ± 0.22

Table 14: Transfer and local-utility results, using the same uncertainty convention. ImageNet target is a percentage; retain and peer losses are percentage points.

Method	ImageNet target	ImageNet retain drop	CIFAR peer drop
GA	46.32 ± 0.52	0.90 ± 0.00	25.79 ± 0.67
NegGrad+	76.01 ± 0.12	**0.11 ± 0.00**	3.47 ± 0.07
SSD	78.37 ± 0.62	1.80 ± 0.06	1.31 ± 0.26
SalUn	24.01 ± 5.97	5.54 ± 0.59	35.97 ± 2.03
Same-knowledge margin	0.07 ± 0.15	4.30 ± 0.06	14.61 ± 0.50
CSOA	0.07 ± 0.15	3.34 ± 0.17	8.92 ± 0.69
Canonical erase	0.13 ± 0.18	3.20 ± 0.18	7.66 ± 0.73
No hierarchy (joint)	20.77 ± 6.09	0.81 ± 0.04	5.78 ± 0.38
No L_{ctrl}	26.01 ± 7.66	0.76 ± 0.05	4.92 ± 0.48
No L_{struct}	**0.00 ± 0.00**	3.39 ± 0.18	10.69 ± 0.73
Broader root	0.20 ± 0.18	3.74 ± 0.23	15.86 ± 0.87
Peer centroid	0.20 ± 0.18	3.32 ± 0.25	2.21 ± 0.68
CLIPErase	0.20 ± 0.08	32.39 ± 0.38	9.13 ± 1.66
CCUP	1.00	1.54	3.43
CSOA-R	0.93 ± 0.43	1.40 ± 0.10	**0.37 ± 0.18**

Table 15: Fixed CSOA by concept, averaged over five seeds. A negative peer drop means improvement. The unmapped ImageNet cells are omitted from transfer averages.

Concept	CIFAR target	Retain drop	ImageNet target	Descriptions	Peer drop
apple	0.00	0.52	0.40	0.00	7.20
bicycle	0.00	0.61	0.00	0.00	5.95
butterfly	0.00	2.50	0.00	0.00	10.10
elephant	0.00	0.83	0.00	0.07	13.60
lawn mower	0.00	0.53	–	0.01	1.30
mountain	0.40	0.36	–	10.87	-1.70
rose	0.00	1.23	–	0.03	23.50
tiger	0.00	1.49	0.00	0.00	0.80
whale	0.00	1.95	0.00	0.00	17.85
woman	0.00	1.00	–	0.00	10.60

two Pets order-level records. Each point compares target images to a normalized text-bank centroid, with the source evaluator averaging batch means. The last point is a mixed reference bank, not one peer breed: the samoyed bank contains great pyrenees, pomeranian, husky, Persian cat, and Abyssinian cat; the Abyssinian bank contains Persian cat, Siamese cat, Bengal cat, beagle, pug, and samoyed. This target-image similarity profile is distinct from exact-label recognition on held-out peer images. The table below retains all twelve source records rather than treating them as new five-seed replications. The original Flowers classifier in these records has very low target and retain accuracy (retain accuracy is around 1% for 102 classes). These Flowers runs therefore cannot establish effective removal from an initially capable classifier; they are included only as historical hierarchy-profile records. The matched ten-concept interventions provide the principal quantitative hierarchy evidence.

Table 16: Original Scene B records. Accuracy columns are percentages. Δ target is the edited-minus-original target cosine; control sim. is edited cosine to the selected control. These single historical runs are not pooled into the multi-seed aggregates. Bold compares edited target and retain accuracies across controls for the same target.

Record	Target accuracy Original	Target accuracy CSOA	Retain accuracy Original	Retain accuracy CSOA	Δ target cosine	Control sim.
abyssinian cat ctrl family	88.78	**0.00**	84.43	**77.96**	-0.12	0.64
abyssinian cat ctrl order	88.78	**0.00**	84.43	77.82	-0.18	0.61
abyssinian cat ctrl species	88.78	**0.00**	84.43	76.03	-0.12	0.64
flower class1 wild ctrl family	0.00	**0.00**	0.95	0.75	0.30	0.56
flower class1 wild ctrl kingdom	0.00	**0.00**	0.95	0.70	0.28	0.59
flower class1 wild ctrl type	0.00	**0.00**	0.95	**0.82**	0.36	0.66
flower class51 garden ctrl family	3.08	**0.00**	0.92	0.69	0.33	0.65
flower class51 garden ctrl kingdom	3.08	**0.00**	0.92	**0.72**	0.25	0.60
flower class51 garden ctrl type	3.08	**0.00**	0.92	0.67	0.39	0.74
samoyed dog ctrl family	100.00	**0.00**	84.11	**78.06**	-0.01	0.62
samoyed dog ctrl order	100.00	**0.00**	84.11	76.60	-0.12	0.62
samoyed dog ctrl species	100.00	**0.00**	84.11	77.28	-0.01	0.63

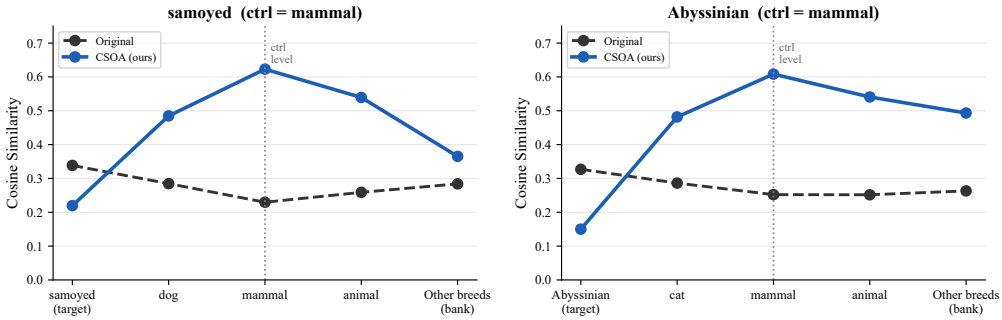

Figure 5: Hierarchy profiles from Scene B: samoyed (left) and Abyssinian (right), both with control mammal. CSOA lowers target similarity and raises control similarity. Similarity to the other-breed reference bank also changes, motivating separate exact-label utility tests. These historical single-run profiles are not pooled into the five-seed averages.

C.5 Cross-style protocol and image counts

The new frozen-checkpoint evaluation uses six targets: apple, bicycle, butterfly, elephant, tiger, and whale. There is no DomainNet adaptation or selection by test performance. All official test images of a predeclared 26-class pool are included: 4,378 real, 1,940 sketch, and 1,144 clipart images. Target image counts are given below; all other classes in this pool form the retain set for each request. Predictions use either all 345 DomainNet names or the shared CIFAR-100 gallery. Thus the native gallery has 345 classes, while the retain-image evaluation covers 26 classes, not the full DomainNet distribution.

The fixed-CSOA records, all factorial variants, and the competing frozen checkpoints use the same test indices and galleries; saved prediction hashes and source-state identities link each record to its source edit. The ten-concept factorial study contains 400 source units; its six-concept frozen tests

Table 17: Per-target image counts in the frozen DomainNet test. Each image set is evaluated for all five source-training seeds.

Concept	Real	Sketch	Clipart
Apple	17	55	27
Bicycle	211	103	22
Butterfly	195	75	48
Elephant	234	80	35
Tiger	179	115	94
Whale	200	80	103

and the CCUP/CLIPErase comparisons contain 828 domain units. SLUG and Text-image Space add 180 domain units. Repeated seeds share test images and are not independent datasets. With the shared gallery, original target accuracies are 96.04%, 89.53%, and 86.59% for real, sketch, and clipart; edited accuracies are 0%, 0%, and 0.37%. The clipart originally-correct conditional residual and normalized BGG are both 0.48%/points after macro averaging, though their definitions differ. The native-gallery zero residual is therefore reported together with the small shared-gallery residual. Earlier forget-coverage experiments and fixed-coverage inference-scaling diagnostics used different training variants; their results are not merged into this frozen comparison or used as a causal proof of the anchor-based direction construction.

Table 18: Frozen target residuals (%) with the same CIFAR-100 gallery in all styles. The original row is a reference; bold marks the lowest edited-model residual. These values are distinct from the native-gallery columns of Table 4.

Method	Real↓	Sketch↓	Clipart↓
Original CLIP	96.04	89.53	86.59
CLIPErase	**0.00**	**0.00**	**0.00**
CCUP	**0.00**	**0.00**	**0.00**
SLUG	96.04	89.53	86.59
Text-image Space	8.64	10.61	25.86
CSOA	**0.00**	**0.00**	0.37
CSOA-R	0.07	0.13	1.71

C.6 Directional penalty and routing

Directional regularization. A separate five-concept, five-seed ablation removes L_o while preserving the remaining CSOA recipe. Across the 25 paired runs, mean changes (no-penalty minus CSOA) in CIFAR target accuracy, ImageNet target accuracy, and description accuracy are all zero. Mean retain-drop changes are -0.000404 points on CIFAR and $+0.000160$ points on ImageNet. This limits the claim of necessity for this penalty; it does not remove the anchor construction, all projections, or the gradient-conflict operation simultaneously. The isolated control, structure, and gradient-projection interventions provide complementary evidence for semantic redirection and selective preservation.

Optional density-aware routing. The diagnostic router gates a frozen adapter using source-model image embedding z:

$$\gamma = \sigma\left(\frac{\max_i \langle z, t_{\text{erase},i}\rangle - \theta}{\tau}\right), \qquad y = Wh(x) + \gamma W_B W_A h(x), \qquad \hat{z} = \text{norm}(y). \quad (7)$$

Here θ is calibrated from retain similarities, τ controls sharpness, W is the frozen visual projection, and $W_B W_A$ is the rank-r adapter. Main results fix $\gamma = 1$ throughout training and evaluation. Routing is a separate utility–coverage sensitivity analysis, not part of the primary CSOA claim. The routing analysis freezes a text-maximum scorer and uses a retain calibration quantile corresponding to $\alpha = 0.05$. The main adapter checkpoint and strength remain unchanged. For ten concepts and five seeds, CIFAR target residual becomes 5.90%, CIFAR retain loss 0.296 points, ImageNet target residual 4.967%, and ImageNet retain loss 0.245 points. Sensitivity settings $\alpha = 0.02$ and 0.10 yield

CIFAR residuals of 11.68% and 3.88%, respectively, with retain losses 0.166 and 0.393 points. They are sensitivity points rather than replacement primary settings chosen after viewing test results. The additional five concepts not used to choose the global router have higher source residual than the five used for router selection (10.36% versus 1.44%); they were nevertheless present in other development experiments. This result illustrates the difference between source retain-FPR control and transferable target recall.

Efficiency. The rank-16 adapter contains 20,480 trainable parameters. Cache-based optimizer-loop timing excludes feature extraction, model loading, and evaluation, and is not an end-to-end speed comparison against full-model baselines.

D VALIDATION PROTOCOL AND CSOA VARIANTS

D.1 VALIDATION AND FIXED OPERATING POINTS

The reported CSOA variants use one recipe and edit strength shared across all concepts. Validation uses the 100 forget and 1,000 retain images reserved from CIFAR training data, with QAS training-anchor queries. Neither the official test images, ImageNet scores, DomainNet scores, nor the external-description test bank participates in selection.

We first require mean target residual $\leq 1\%$, mean anchor-query residual $\leq 5\%$, and HCS $\geq 90\%$. Among feasible candidates, selection minimizes

$$\max(0, \Delta A_r) + 0.25 \max(0, \Delta A_{\text{peer}}) + 0.05 A_{\text{anchor}}.$$

Ties use lower target residual, higher HCS, lower strength, and then the recipe name. The same validation rule ranks the selected variants across families. The operating points are fixed using validation on seeds 0–2, then evaluated with seeds 0–4. Earlier experiments informed configuration development; the concepts are not claimed as previously unseen development targets.

Table 19: Validation-selected family operating points. Residuals and HCS are percentages; retain and peer losses are points. Bold marks the best displayed value in each performance column. Strength is a setting, not a performance measure.

Family	Strength	Target	Anchors	HCS	Retain loss	Peer loss
A: Loss weights	0.75	0.63	2.99	90.87	0.03	1.80
B: Rank and learning rate	0.75	0.67	2.58	91.30	0.02	2.44
C: CSOA-R	1.00	0.27	1.27	**94.33**	-0.01	0.20
D: Control margin	1.00	**0.00**	1.16	91.10	-0.02	4.59
E: Erase margin	0.75	0.07	**0.32**	92.27	**-0.04**	**0.00**
F: Initialization	0.50	0.43	2.80	92.70	0.07	2.62

D.2 SELECTED VARIANT DEFINITIONS

All variants retain the same source partitions and semantic resources. The following are the fixed settings used for the reported results; different columns do not select different configurations.

Retention-focused CSOA (C, CSOA-R). The retain objective adds KL divergence over all 100 canonical class texts at temperature 0.07, with coefficient 4. One quarter of each retain batch is drawn from same-coarse peer classes within the original retain training pool. The control coefficient is 1.5, compared with 3 for fixed CSOA. The retain coefficient is 6; rank 16, ten epochs, learning rate 3×10^{-4}, and strength one remain unchanged. The KL term enters the protected retain gradient. These three changes define one combined recipe. The full factorial experiment below isolates each change at the same source partitions, initialization convention, training budget, final epoch, and edit strength.

Other reported operating points. Variant A uses control/retain coefficients 1.5/24 and strength 0.75. Variant B uses rank 16, learning rate 3×10^{-4}, ten epochs, control/retain coefficients 1.5/12, and

strength 0.75. Variant D replaces control alignment by $\mathbb{E}_f[\max(m + \max_{t \in T_{\text{cls}} \cup T_e} \hat{z}^\top t - \hat{z}^\top u_c, 0)]$, with margin 0.08, control/retain coefficients 3/12, and strength one. Variant E combines a bounded erase term $\mathbb{E}_f[\max(m + \max(\hat{z}^\top t_c, \max_{t \in T_e} \hat{z}^\top t) - \max_{j \neq c} \hat{z}^\top t_j, 0)]$ with bounded control, margin 0.03, control/retain coefficients 1.5/12, no added class KL, and strength 0.75. It ranks first under the source-validation rule. Variant F uses the random-initialization control at rank 32, learning rate 10^{-4}, control/retain coefficients 1.5/12, and strength 0.5; no benefit is attributed to ridge initialization.

D.3 Comparative operating points

Table 20: All six selected families and fixed CSOA on the official test sets, five seeds. CIFAR and ImageNet target residuals, descriptions, and HCS are percentages; utility losses are points. Bold follows each metric's direction. Family E is the global source-validation winner.

	CIFAR		ImageNet		Desc.	HCS	Peer
Configuration	Target	Ret. loss	Target	Ret. loss			loss
CSOA	**0.04**	1.10	**0.07**	3.34	**1.10**	**98.72**	8.92
A: Loss weights	1.28	0.02	1.67	0.44	4.25	89.76	1.89
B: Rank and learning rate	1.04	0.04	1.37	0.52	3.89	90.52	2.48
C: CSOA-R	0.22	0.08	0.93	1.40	2.60	94.98	**0.37**
D: Control margin	0.06	**-0.05**	0.55	1.17	16.07	90.80	4.78
E: Erase margin	0.10	0.01	15.90	**0.15**	6.40	92.10	0.47
F: Initialization	0.68	0.07	1.61	0.32	4.57	92.54	2.21

The fixed recipe has the lowest target residual on both CIFAR and ImageNet and the lowest description residual among these seven CSOA configurations. CSOA-R provides a separate retention-oriented operating point, reducing global and neighboring-class damage with a modest increase in residual recognition. Family E's low source-validation cost does not transfer to low ImageNet residual. We retain all family outcomes, including this global-selection result, rather than substituting the family with the most favorable test average. CSOA-R is reported as a named prespecified-family variant, not as the global validation winner or as a uniform replacement for the fixed recipe. The independent CSOA-R factorial and frozen-transfer tests below close this gap without substituting fixed-CSOA measurements for CSOA-R.

Table 21: CSOA-R by concept, averaged over five seeds. Missing ImageNet mappings are omitted from the transfer average. Columns use the same units as the preceding table.

	CIFAR		ImageNet		Desc.	HCS	Peer
Concept	Target	Ret. loss	Target	Ret. loss			loss
apple	0.00	0.02	2.00	0.17	0.80	96.20	-2.30
bicycle	0.00	-0.06	0.00	0.80	0.00	96.60	0.30
butterfly	0.00	0.34	0.00	2.44	0.00	98.20	2.00
elephant	0.00	-0.02	0.00	0.56	1.41	90.00	1.40
lawn mower	0.00	-0.42	–	–	0.36	93.80	-6.55
mountain	1.00	0.03	–	–	21.99	95.00	-1.00
rose	0.00	0.01	–	–	0.40	95.20	-1.00
tiger	0.00	0.02	0.00	2.94	0.01	89.80	0.15
whale	1.20	0.75	3.60	1.49	0.52	97.20	5.40
woman	0.00	0.14	–	–	0.49	97.80	5.30

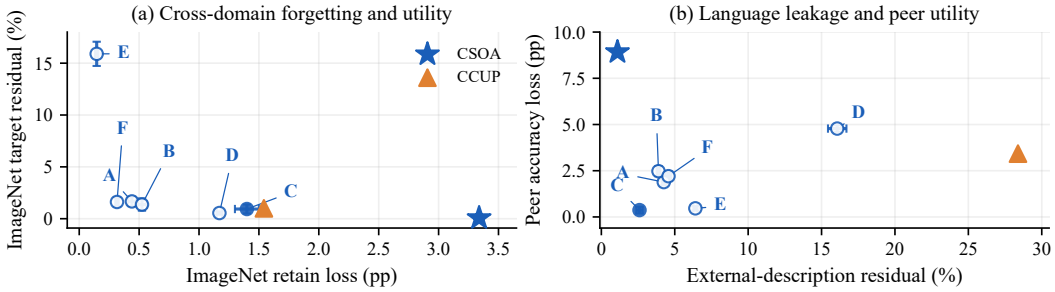

Figure 6: Test trade-offs of all six validation-selected CSOA families: A, loss weights; B, rank/learning rate; C, CSOA-R; D, control margin; E, erase margin; F, initialization. Error bars show sample SD across five seed-level means. Lower is better on both axes. The star is fixed CSOA; the triangle is CCUP.

Factorial attribution of CSOA-R. We cross class-distribution distillation (weight zero or four), same-coarse peer replay (fraction zero or 0.25), and control weight (three or 1.5). All eight combinations are evaluated on ten concepts and five seeds, with no per-metric reselection. Table 22 gives the complete source/transfer outcomes. Figure 2 averages each on-minus-off contrast over the four settings of the other two factors within concept and seed. Circles average over seeds; error bars on the macro means are the sample SD of five seed-level concept averages. These are average effects at the tested settings, not independent additive contributions to the full-versus-default difference.

Table 22: Complete CSOA-R factorial study, ten concepts and five seeds; ImageNet uses six mapped concepts. Target/description accuracy and HCS are percentages; retain/peer losses are points. Bold follows each metric's direction.

	CIFAR		ImageNet		Desc.	HCS	Peer
Configuration	Target	Ret. loss	Target	Ret. loss			loss
CSOA	**0.04**	1.10	**0.07**	3.34	**1.10**	98.72	8.92
Lower control only	**0.04**	0.24	**0.07**	1.12	2.46	94.16	5.03
KD only	0.08	0.59	0.13	2.99	1.26	**98.76**	5.09
Peer replay only	0.14	1.16	0.13	3.64	1.18	98.60	3.07
CSOA-R: restore control	0.34	0.70	0.13	3.38	1.35	98.60	1.61
CSOA-R: no KD	0.08	0.44	0.13	1.77	1.81	94.64	2.09
CSOA-R: no peer replay	0.16	**-0.01**	0.67	**0.85**	3.34	94.78	1.88
CSOA-R	0.22	0.08	0.93	1.40	2.60	94.98	**0.37**

The averaged effects expose complementary but non-monotonic roles. Distillation reduces overall retain loss by 0.40 points and peer loss by 2.54 points, while increasing description residual by 0.50 points. Peer replay reduces peer loss by 3.45 points but increases overall retain loss by 0.11 points. Lower control weight reduces overall retain loss by 0.70 points while increasing description residual by 1.33 points. The complete combination gives lower peer loss than changing control weight alone; its value is selective preservation rather than uniform improvement on every measure.

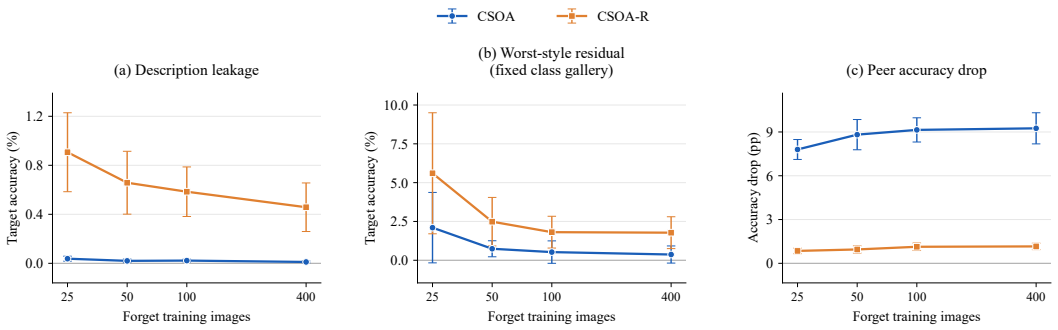

Figure 7: Limited-data forgetting with a fixed optimization budget, six concepts and five seeds. Worst-style residual takes the maximum over real/sketch/clipart within each concept and seed using the same CIFAR gallery, then macro-averages. Error bars are SD across seed-level means. CSOA-R preserves peers better, while CSOA has lower residual recognition.

Few-shot source adaptation and frozen transfer. The six mapped concepts use 25, 50, 100, or 400 forget images and five seeds for each of CSOA and CSOA-R. Subsets are nested within each saved source-training split; validation and official test images remain excluded. Ten epochs and 32 optimization steps per epoch are held fixed, with sampling that avoids dropping a small forget set. The five context anchors use the first 16 images inside the selected subset. The 400-image reference reuses completed checkpoints. Each source checkpoint is evaluated on real, sketch, and clipart without adaptation, yielding 240 source and 720 frozen-domain units.

Table 23: Limited-data results on the same six concepts. Parentheses give forget training images. Target/description residuals are percentages; utility losses are points. DomainNet uses its native gallery here; Figure 7 instead reports worst-style residual with the fixed CIFAR gallery. Bold compares methods at the same sample budget.

		Source			Frozen DomainNet target		
Method (images)	Target	Ret. loss	Desc.	Peer loss	Real	Sketch	Clipart
CSOA (25)	**0.00**	1.16	**0.04**	7.80	**0.04**	**0.27**	**0.21**
CSOA (50)	**0.00**	1.31	**0.02**	8.82	**0.00**	**0.00**	**0.00**
CSOA (100)	**0.00**	1.34	**0.02**	9.14	**0.00**	**0.00**	**0.00**
CSOA (400)	**0.00**	1.32	**0.01**	9.25	**0.00**	**0.00**	**0.00**
CSOA-R (25)	0.60	**0.18**	0.91	**0.85**	1.03	1.06	1.61
CSOA-R (50)	0.37	**0.19**	0.66	**0.95**	0.12	0.25	0.93
CSOA-R (100)	0.20	**0.20**	0.58	**1.13**	0.03	**0.00**	0.49
CSOA-R (400)	0.20	**0.17**	0.46	**1.16**	**0.00**	0.04	0.37

Curves first average concepts within each seed, then average the five seed means. The worst-style statistic first maximizes real/sketch/clipart target accuracy within each concept and seed, using the same 100-class gallery, and only then performs macro averaging. It is not the maximum of three domain-level macro means. Seeds share the same test images. At 25 images, the largest individual CSOA-R source residual is 9% for whale, seed one; its mean of 0.60% must not be interpreted as uniform removal.

E GENERATIVE-MODEL EVALUATION

Architecture and intervention. The generative diagnostic experiments freeze Qwen2.5-VL-7B-Instruct and insert a rank-16 residual after its visual merger. The edit has the form $v' = v + s\,a\,g(v)BAv$, where a is the saved coverage multiplier and g is either an image-conditioned gate or one. The gate is computed from original visual features; training and scoring do not update the language-model backbone. This is an adaptation of the CSOA design to a different architecture, with its own loss and calibration protocol, rather than the identical CLIP objective. Greedy generation

uses fixed prompts. Each fresh comparison regenerates all conditions in one environment; historical answer strings are not assumed to remain identical across runtime changes.

Pets. The four target breeds are pug, samoyed, Sphynx, and shiba inu, each evaluated with five seeds and two routing modes (40 checkpoints). The replay uses the saved adaptation splits and separate official trainval calibration/validation images; validation includes 20 target and 20 matched images. The strength selection is frozen before the enlarged test replay, and the same selected checkpoint/strength is used for the additional query and generation measurements. Selection matches target-positive Yes and balanced-accuracy outcomes using a frozen validation rule, with smaller strength breaking ties; this is not equal free-form performance or equal utility by construction. The replay checks checkpoint identity, separation of adaptation/calibration/validation/test filenames, restoration of base logits, equality of zero-strength generation, and the base-model sentinel after evaluation. Generation prompts ask for a specific breed, an open description, or a response that permits a coarse category. Table 24 separates the fixed breed-identification output from positive-response and hierarchy tests.

Table 24: Pets at frozen validation operating points. All entries are percentages except the global non-target utility drop (points). A low positive Yes rate need not imply low free-form disclosure.

Mode	Target Yes	Held-out alias Yes	Free-form target	Parent BA	Utility drop
Ungated	**3.02**	**6.31**	91.83	**100.00**	**-0.11**
Gated	10.00	10.13	**18.83**	**100.00**	0.33

Across three generation prompts, the target lexical-match means are 77.28% ungated and 16.06% gated; these differ from the specific breed-identification column because the query distribution differs. The 7,200 saved semantic-generation records are evaluated by fixed lexical rules; human semantic annotation is not part of this evaluation. Consequently, lexical absence is not counted as proof of correct ancestor generation, and parent balanced accuracy is not a substitute for this distinct assessment. The general non-target utility average and matched-class fine-grained generation are also distinct measurements.

DomainNet. The eight requested concepts are apple, banana, bicycle, bus, butterfly, elephant, tiger, and whale. For each of three seeds and two routing modes, adaptation uses 20 source-real target images and 80 retain images (20 from each of four retain classes). Five calibration and five validation images per class are disjoint from adaptation. Each real/sketch/clipart test class uses at most 20 images without replacement, with a minimum of five. The residual is trained for ten epochs at learning rate 5×10^{-4}, retain weight 8, control weight 2, and locality weight 10^{-3}. The source retain-FPR calibration target is 0.2; rank is 16; the coverage multiplier is 0.5; evaluation strength is one. These settings are not pooled with the Pets strength-selection protocol.

The original-model validation criterion is target balanced accuracy $\geq 70\%$ and category-name lexical generation accuracy $\geq 50\%$. All eight concepts are run and retained in the artifacts; apple fails the capability screen (40% validation free-form accuracy), leaving seven concepts in the primary macro average. The test data are not used for this inclusion decision. The original free-form accuracies for the included concepts are 93.57% real, 95.00% sketch, and 88.10% clipart.

Table 25: Historical DomainNet diagnostics across seven validation-qualified concepts and three seeds. Target and alias residuals and gate recall are percentages; retain utility drops are points. Ungated recall is 100% by construction.

Mode	Domain	Free-form target	Alias Yes	Gate recall	Retain drop
Ungated	real	**2.14**	**22.62**	**100.00**	2.38
Ungated	sketch	**2.86**	**20.71**	**100.00**	2.92
Ungated	clipart	**5.00**	**19.05**	**100.00**	2.17
Gated	real	22.38	39.05	77.14	**1.13**
Gated	sketch	66.67	75.00	29.29	**0.15**
Gated	clipart	61.90	73.81	28.81	**0.12**

These generative experiments test evaluation behavior using a different adaptation and calibration protocol from the primary CLIP study. They do not establish a general generative-VLM unlearning framework or transfer of the CSOA-R recipe. They are supporting evidence of the distinction between recognition suppression and answer disclosure, not a pooled estimate of CLIP performance.

Paired Qwen generation. All eight requested concepts, three seeds, and three styles are regenerated under four conditions: original, saved gated CSOA, the same checkpoint with its gate forced on, and separately trained ungated CSOA. The primary table retains the seven concepts qualified by the historical source-validation criterion; the complete eight-concept table is also shown. Every condition uses paired images and prompts in the same runtime. The matched column measures fine-class generation on non-target images, not binary accuracy or overall utility. These fresh answers are not pooled with the historical binary/lexical records above. Turning the gate on for the same checkpoint

Table 26: Paired Qwen generation for seven validation-qualified concepts and three seeds. Each domain shows target lexical residual (T, %, lower is better) and matched-class lexical correctness (M, %, higher is better) under the specific-category prompt. Bold compares edited conditions.

	Real		Sketch		Clipart	
Condition	T↓	M↑	T↓	M↑	T↓	M↑
Original	93.57	85.08	95.71	76.43	88.10	75.55
CSOA (gated)	22.62	**71.27**	67.38	**72.86**	61.90	**76.03**
CSOA (gate fixed on)	2.62	40.24	5.24	38.57	8.89	48.91
CSOA (ungated training)	**2.14**	54.03	**2.86**	52.86	**5.00**	57.92

Table 27: The same Qwen protocol including all eight requested concepts. Apple is retained here despite failing the historical validation-capability criterion. Units and bolding follow Table 26.

	Real		Sketch		Clipart	
Condition	T↓	M↑	T↓	M↑	T↓	M↑
Original	90.70	86.32	91.88	78.75	85.21	74.44
CSOA (gated)	19.79	**74.24**	65.21	**75.63**	59.79	**74.86**
CSOA (gate fixed on)	2.29	47.08	4.58	45.21	7.78	51.13
CSOA (ungated training)	**1.88**	59.15	**2.50**	57.92	**4.38**	59.01

isolates an inference intervention: less target disclosure comes with lower matched-class correctness. Separately training without a gate recovers some utility relative to post-hoc gate removal. These findings distinguish coverage failure from a successful semantic abstraction and do not identify the CLIP CSOA-R mechanism in a generative model. CSOA-R-VLM improves both direct-category

Table 28: Paired Qwen comparison on eight predeclared concepts, three seeds, and three DomainNet styles. T is direct-category target lexical disclosure (%, lower is better); M is matched-class naming correctness (%, higher is better). CSOA and CSOA-R-VLM checkpoints are chosen by source validation. Every row uses the same images and prompts within its concept–seed unit. Bold marks the best edited value in each column.

	Real		Sketch		Clipart	
Condition	T↓	M↑	T↓	M↑	T↓	M↑
Original	90.70	86.32	91.88	78.75	85.21	74.44
CSOA	**3.75**	76.53	15.83	67.50	14.38	65.90
CSOA-R-VLM	5.00	**84.29**	**5.42**	**75.83**	**7.57**	**75.06**
GA+KL	10.00	75.50	10.42	66.88	9.86	70.99
NPO+KL	10.00	81.91	10.83	71.67	11.74	74.16

target suppression and matched-class correctness over GA+KL and NPO+KL in all three styles. This statement concerns the paired direct-category prompt, not all generation tasks: its open-description target disclosure is 9.17/13.75/23.54% on real/sketch/clipart, while GA+KL gives 13.33/9.38/11.32%. On clipart, other retained-class naming is 69.37% for CSOA-R-VLM versus 70.31% for GA+KL and

71.80% for NPO+KL. CSOA-R-VLM has a different training budget from the adapted comparison objectives, and all scores are lexical, not human semantic judgments.

LLaVA comparison and retention recovery. We evaluate LLaVA-1.5-7B (Liu et al., 2024) on pug and samoyed with five seeds. A trainable visual residual edits frozen visual and language backbones. Nine conditions generate paired answers to specific, open, and coarse-permitted prompts, which are distinct query distributions.

Source validation compares gated and always-on CSOA at strengths 0.5/1/1.5. CSOA-R-VLM continues its checkpoint for three epochs at learning rate 10^{-4} and retain weight eight, adding answer-token KL and 50% matched-peer replay while testing control/KL weights 1/1 or 0.5/4. This adds training budget and differs from CLIP's class-level KL. Selection uses five validation images per class and two target prompts. Feasible candidates have lexical target residual $\leq 10\%$ and retain drop ≤ 10 points; otherwise the least violation is flagged. Seven of ten units meet the constraints for each selected variant; all ten enter test aggregates.

Table 29: Paired LLaVA comparison on two breeds and five seeds. Target columns give lexical residual (%, lower is better) for each prompt; matched and other-retain columns give specific-name lexical correctness (%, higher is better). CSOA denotes the source-validation choice among gate and strength settings. Bold compares edited conditions at displayed precision. These lexical scores do not distinguish appropriate abstraction from incorrect names or refusal.

	Target residual↓			Retained correctness↑	
Condition	Specific	Open	Coarse	Matched	Other
Original	55.00	32.50	47.50	55.00	51.88
CSOA (gated)	32.50	15.00	37.00	55.00	51.88
CSOA (gate fixed on)	**0.00**	**0.00**	1.50	47.50	39.63
CSOA	2.50	**0.00**	11.00	47.00	40.00
CSOA-R-VLM	2.50	**0.00**	24.50	55.00	51.88
GA+KL	**0.00**	**0.00**	**0.00**	**64.00**	51.38
NPO+KL	**0.00**	**0.00**	**0.00**	55.50	**52.13**
SimNPO+KL	55.00	32.00	47.50	55.00	51.63
KVW	47.50	25.00	35.50	57.50	51.63

CSOA-R-VLM restores original matched/other-retain correctness, but coarse-prompt residual rises to 24.50%, versus 11% for selected CSOA. This prompt sensitivity motivates the separate follow-up below; GA+KL and NPO+KL remain competitive.

Generative comparison implementations. GA+KL, NPO+KL (Zhang et al., 2024), and SimNPO+KL (Fan et al., 2025) optimize answer-token forgetting with retained-answer KL under the same visual-residual parameterization. They are objective-function transfers to this Pets protocol, not replications of the original language-only benchmark settings. KVW (Kim et al., 2026) instead edits decoder down-projection columns using retained-answer activation statistics, one pass through forget samples, coefficient 0.02, and layers 1–25. The trainable modules and optimization budgets therefore differ; the table is not a matched-budget Pareto comparison.

Table 30: Independent FP32 LLaVA comparison: two breeds, five seeds, paired test images and prompts. Target columns are lexical residuals (%, lower is better); matched and other columns are exact-name correctness (%, higher is better). The MMUnlearner objective is adapted to this protocol, not a replication of its original benchmark. Absolute values are not pooled with the separate FP16 nine-condition study.

	Target residual↓			Retained correctness↑	
Condition	Specific	Open	Coarse	Matched	Other
Original	55.00	32.50	47.50	55.00	51.88
CSOA	2.50	0.00	11.00	47.00	40.00
CSOA-R-VLM	2.50	0.00	24.50	55.00	52.00
MMUnlearner	52.50	32.00	46.50	55.00	52.62

In this FP32 comparison, the adapted MMUnlearner (Huo et al., 2025) objective retains the original model's 55.00% matched accuracy but leaves 52.50% direct-answer target residual. CSOA-R-VLM has 2.50% residual at the same matched accuracy, while CSOA has lower coarse-prompt residual but greater retained-class damage. This supports a useful operating point under the adapted protocol, not a claim about the original MMUnlearner benchmark.

Table 31: LLaVA comparison on the original test panel (top) and disjoint test images (bottom), two breeds and five seeds. Values are percentages. Target columns measure lexical residual; retained columns measure specific-name correctness. Bold compares edited conditions within each panel. The parent-loss row adds parent-answer supervision; the multi-prompt row uses expanded validation without this loss.

	Target residual↓			Retained correctness↑	
Condition	Specific	Open	Coarse	Matched	Other
Original test panel					
Original	55.00	32.50	47.50	55.00	51.88
CSOA	17.50	9.00	23.00	51.50	47.50
CSOA-R-VLM	2.50	**0.00**	24.50	55.00	51.88
CSOA-R-VLM (multi-prompt)	2.50	**0.00**	22.50	55.00	52.00
CSOA-R-VLM + parent loss	**0.00**	**0.00**	**0.00**	55.00	50.75
GA+KL	**0.00**	**0.00**	**0.00**	**64.00**	51.38
NPO+KL	**0.00**	**0.00**	**0.00**	55.50	**52.13**
Additional disjoint test images					
Original	57.50	37.50	42.50	60.00	54.38
CSOA	15.50	10.00	15.50	57.00	49.50
CSOA-R-VLM	2.50	**0.00**	14.00	60.00	53.75
CSOA-R-VLM (multi-prompt)	2.50	**0.00**	14.00	60.00	53.75
CSOA-R-VLM + parent loss	**0.00**	**0.00**	**0.00**	60.00	52.25
GA+KL	**0.00**	**0.00**	**0.00**	**66.50**	53.13
NPO+KL	**0.00**	**0.00**	**0.00**	57.00	**54.00**

Generation-aligned retention refinement. The follow-up retains both breeds and five seeds, adding 20 official-test images per class disjoint from prior panels, training, and validation. Seven conditions are freshly paired in each panel; this tests new images of the same concepts.

CSOA-R-VLM refinement fixes answer-token KL weight four, matched-peer replay fraction 0.5, learning rate 10^{-4}, and retain weight eight. It crosses control weights 0.5/1 with generation-supervision coefficients 0/0.5. Training runs for six epochs, with candidate checkpoints at epochs three and six and strengths 0.75/1/1.5. The generation-supervised variant additionally minimizes

$$L_{\text{gen}} = \mathbb{E}_{(x,p)\sim\mathcal{D}_f}\big[-\log P_{\theta+\Delta}(a_{\text{parent}} \mid x, p)\big], \tag{8}$$

where a_{parent} is the declared permitted parent answer and p cycles through the specific, open, and coarse-permitted training prompts. Thus this variant changes the training objective, rather than only selecting a better checkpoint of the original CSOA-R-VLM recipe.

Validation caps the worst of three prompt residuals at 10% and each retained class drop at ten points. Feasible candidates prioritize utility; otherwise, the least-violating candidate is retained. CSOA, multi-prompt CSOA-R-VLM, and its parent-supervised extension are feasible in 2/10, 5/10, and 7/10 units; all ten are reported. The extension uses control/generation weights 0.5/0.5, epoch three, and strength 0.75.

Pure refinement leaves 14% coarse-prompt residual on new images; parent-answer supervision yields zero lexical residual across prompts in both panels. New-image matched correctness remains 60%, with other-retain correctness changing from 54.38% to 52.25%. GA+KL and NPO+KL remain competitive. These results support generation-aware control, without proving dominance or complete information removal.

Fixed lexical matching cannot distinguish a permitted ancestor from a wrong fine class, refusal, or unrelated answer. Human semantic verification remains necessary.