DriftScope: Measuring The Hidden Effects of Diffusion Model Adaptation

Héctor Laria*[1,2], Yiping Han*[1,2], Julian D. Santamaria[1,2], Kai Wang[3,4]†, Bogdan Raducanu[1,2], Joost van de Weijer[1,2], and Alexandra Gomez-Villa[1,2]

[1] Computer Vision Center, Barcelona, Spain
[2] Universitat Autonoma de Barcelona, Barcelona, Spain
[3] Program of Computer Science, City University of Hong Kong (Dongguan), China
[4] City University of Hong Kong, HK SAR, China
{hlaria, yhan, jsantamaria, bogdan, joost, agomezvi}@cvc.uab.es, kai.wang@cityu-dg.edu.cn

Abstract. Adapting pre-trained text-to-image diffusion models, whether to learn new visual concepts or erase unwanted ones, is routinely evaluated on its intended effects alone. We argue this framing is incomplete. Through sparse autoencoder analysis and zero-shot classification, we demonstrate that adaptation systematically damages semantically unrelated concepts in ways that aggregate metrics structurally cannot surface: when damage is severe enough for FID and KID to respond, the model is already nearly unusable; when the model remains functional, FID and KID stay flat while specific classes silently suffer worst-case zero-shot accuracy drops of up to 18.9 points and concept-level distributions shift dramatically. This pattern appears at both ends of the adaptation spectrum (concept customization and concept unlearning), suggesting it is a systematic consequence of weight-level modification rather than an artifact of any particular method. To surface this hidden drift before deployment, we introduce DriftScope, a prompt-level diagnostic tool that takes any two model checkpoints and returns a ranked list of tokens whose visual concepts have shifted most between them. DriftScope optimizes a soft prompt to attribute drift at the token level without requiring access to real data or model internals. The result is an interpretable, concept-level audit that aggregate evaluation cannot provide.

Keywords: Diffusion models · Model evaluation · Model diffing

1 Introduction

Adapting foundation models to specific needs has become a central challenge in modern AI deployment, and weight-level adaptation is among the most widely

* Equal Contribution.
† Corresponding Author.
Project page: https://hyping111.github.io/DriftScope/

adopted solutions. Whether customizing text-to-image models to new visual concepts [26] or removing unsafe content through unlearning [10], practitioners routinely modify pre-trained weights to serve specific needs. These operations are treated as benign optimizations, evaluated mostly on their intended effects: Does the customized model generate the new concept? Does the unlearned model avoid harmful content?

We argue this framing is incomplete. While adaptation achieves its proximal goals, it simultaneously degrades the model in ways current evaluation misses [17]. Standard metrics (such as FID, KID, prompt-specific evaluations) focus on aggregate quality and target-task performance, missing fine-grained distributional shifts. As a result, adapted models may lose entire categories of concepts and exhibit dramatic perceptual shifts on unrelated tasks, all while passing standard benchmarks. These "hidden costs" arise across opposite ends of the adaptation spectrum: concept customization, which *adds* a new concept, and concept unlearning, which *removes* one, suggesting this is a systematic consequence of weight-level adaptation, not an artifact of any particular method.

The core problem is granularity. Standard metrics measure aggregate quality; they are structurally blind to concentrated damage on specific concepts. What practitioners need is not another aggregate score, but a tool that answers a precise question: *Given this prompt, which words have their generated concepts most affected by adaptation?*

In this paper, we first show that this drift is real and systematic through two complementary analyses: sparse autoencoders (SAEs) to surface concept-level shifts in the generative distribution, and zero-shot classification via GDC [25] to confirm these shifts are reflected in the model's latent space. Together they reveal that adaptation inflicts damage on the model's capabilities while leaving aggregate metrics intact, exactly the failure mode that FID-based evaluation cannot surface. Building on this evidence, we introduce DriftScope, a diagnostic tool that takes any prompt and two model checkpoints (a base model and an adapted variant) and returns a word-level drift report identifying where semantic coverage has eroded. DriftScope generates a prompt using the base model and feeds it through both checkpoints simultaneously, maximizing the distance between the cross-attention maps produced by each model. Words whose cross-attention activations diverge most between the two models are identified as the concepts most affected by adaptation, often surfacing non-obvious collateral damage beyond the adapted target.

In summary, this paper makes the following contributions:

- We show that concept customization and concept unlearning, both induce collateral damage on semantically unrelated concepts. Aggregate metrics such as FID and KID either remain flat while damage accumulates silently, or only respond once the model has become nearly unusable, revealing a structural blind spot in standard evaluation.
- We establish this through sparse autoencoder analysis, which surfaces concept-level shifts in the generative distribution, and zero-shot classification, which confirms these shifts in the model's latent space.

– We introduce DriftScope, a prompt-level diagnostic tool that returns a ranked blacklist of tokens whose visual concepts have drifted most between any two checkpoints, via differentiable cross-attention divergence maximization.

2 Related Work

Divergent representations in T2I models. Divergent representations refer to systematic discrepancies between what a generative model is expected to produce and what it actually generates, whether measured against a reference data distribution or against another model checkpoint. Bohacek *et al.* [3] formalize this as *conceptual blindspots*, using sparse autoencoders over DINOv2 features to identify concepts that are suppressed or exaggerated in generated images relative to training data. In a complementary direction, Tong *et al.* [31] mine for *erroneous agreements* in CLIP and use a language model to surface systematic failure patterns that propagate through downstream T2I systems. Dunlap *et al.* [8] shift the lens to cross-model comparison, introducing COMPCON, an evolutionary search algorithm that discovers visual attributes differing between two *independently trained* T2I models. While effective for model selection, COMPCON relies on iterative calls to external VLMs and LLMs and is non-differentiable by design; because it compares models that differ in architecture, training data, and procedure simultaneously, the divergences it surfaces cannot be attributed to any specific cause. DriftScope targets a more precise setting (both checkpoints share the same architecture and weights, differing only in a targeted adaptation) and is entirely self-contained, using gradient backpropagation through two frozen checkpoints to produce token-level drift scores directly actionable for pre-deployment auditing.

Adversarial Prompting in Text-to-Image Models. A line of work has studied how carefully crafted prompts can expose weaknesses in T2I diffusion models. Liu *et al.* [19] propose SAGE, an adversarial search method that systematically explores the prompt and latent space to surface failure modes. Tsai *et al.* [32] introduce Ring-A-Bell, a model-agnostic red-teaming scheme that constructs adversarial prompts to bypass concept removal mechanisms. Both approaches optimize prompts to probe a *single* model's behavior, either to find failures or circumvent erasure. DEXTER [6] uses differentiable prompt optimization (a masked LM with Gumbel-Softmax relaxation) to maximize neuron activations in a visual classifier, relying on external VLMs for global explanations. DriftScope shares the same prompt construction infrastructure but uses it as a self-contained *measurement instrument* to compare two generative checkpoints, localizing tokens whose visual concepts have shifted as a result of adaptation.

Continual Learning and Catastrophic Forgetting in Generative Models. Catastrophic forgetting [23] has been studied in text-to-image models primarily through the lens of continual customization. Methods such as C-LoRA [28]

and FL2T [14] address the sequential setting where concepts are added one after another; Zając *et al.* [35] further benchmark continual learning baselines within diffusion models, identifying timestep-dependent forgetting patterns. These works share the assumption that forgetting is a *known problem* to be mitigated across a sequence of adaptation steps. Our work addresses a more fundamental prerequisite: in a *single* adaptation step, DriftScope does not prevent forgetting. It surfaces it, providing a concept-level audit that is a necessary precondition for any mitigation strategy. This builds directly on prior work [17], which first characterized open-world forgetting in diffusion model customization via zero-shot classification and appearance drift analysis; here we extend that investigation to concept unlearning and introduce a diagnostic tool to identify *which* concepts are affected.

Model diffing. Model diffing refers to systematically comparing two model checkpoints to localize what has changed between them. Lindsey *et al.* [18] introduce crosscoders, sparse autoencoders trained simultaneously on residual stream activations of two language model checkpoints, producing features that cluster into shared, base-model-specific, and fine-tuned-model-specific groups. Bricken *et al.* [5] tracks similar changes by fine-tuning an existing SAE dictionary on the post-adaptation model, measuring how individual features rotate or are suppressed. Both methods operate over internal activations of language models. DriftScope shares the diagnostic intent but targets a different modality: rather than probing residual stream features, it compares cross-attention maps in text-to-image diffusion models and surfaces concept-level drift at the token level, requiring no access to model internals beyond standard attention outputs.

3 Method

3.1 Preliminaries

Diffusion models. Diffusion models [13,29,30] are a class of generative models that learn to synthesize data by reversing a gradual noising process. The forward process progressively adds Gaussian noise to data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ over T timesteps according to a variance schedule $\{\beta_t\}_{t=1}^{T}$, producing increasingly noisy samples $\mathbf{x}_t$ until $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse process learns to denoise these samples by training a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ to predict the noise added at each timestep, optimizing the variational lower bound: $\mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right]$. At inference time, samples are generated by iteratively denoising from pure noise $\mathbf{x}_T$ through the learned reverse process.

Diffusion model adaptation. We study two paradigms through which practitioners routinely modify pre-trained text-to-image diffusion models.

- **Concept unlearning** aims to surgically remove a target concept from a model's generative distribution without retraining from scratch. We evaluate four methods: Ablating Concepts (AC) [10], which fine-tunes the model

to redirect the target concept toward an anchor; EraseDiff [34], which formulates unlearning as a constrained optimization over the denoising objective; MACE [21], which uses closed-form weight updates via a multi-concept erasure framework; and SPM [22], which trains lightweight concept-specific modules to suppress target activations.
– **Model customization** adapts a pre-trained model to faithfully generate a specific user-provided visual concept from only a handful of reference images. We evaluate four methods: DreamBooth [26], which fine-tunes the full model with a class-preservation loss; Custom Diffusion [16], which restricts fine-tuning to cross-attention key and value projections; BOFT [20], which reparameterizes weight updates via butterfly orthogonal transformations; and SVDiff [11], which fine-tunes singular values of weight matrices. We evaluate across three base architectures (SD 1.5, SD 2.1, and SD 3.5) to assess whether the observed drift patterns are model-dependent.

Throughout this paper, we refer to the **base model** as the original pre-trained text-to-image diffusion model (*e.g.*, Stable Diffusion 3.5), and the **modified model** as the same architecture after adaptation. In all cases, we compare generations from the base and modified models using identical prompts and random seeds to isolate the effects of the adaptation.

3.2 Model Adaptation Induces Hidden Distributional Drift

Concept-Level Drift To probe whether model adaptation alters the generative distribution beyond its intended target, we analyze concept-level shifts using the pipeline introduced by Bohacek *et al.* [3]. Their method trains a sparse autoencoder (SAE) over DINOv2 image features to decompose image embeddings into approximately interpretable concept activations, enabling quantitative comparison of concept prevalence across image sets. Originally designed to compare model-generated images against real training data, we repurpose this diagnostic for a different setting: comparing two sets of images generated by the *same architecture* under two different weight configurations. This eliminates the need for real reference images and allows drift to be attributed directly to the adaptation.

We sample 10,000 prompts from DiffusionDB and generate paired images from the base model $\mathcal{M}$ and its adapted counterpart $\mathcal{M}'$ using identical prompts and noise seeds. Let $\varepsilon_k(x)$ denote the SAE activation corresponding to concept k. For each concept, we compute a drift score:

$$\omega(k) = \sigma(\mathbb{E}_{x' \sim \mathcal{M}'}[\varepsilon_k(x')] - \mathbb{E}_{x \sim \mathcal{M}}[\varepsilon_k(x)]), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $\omega(k) = 0.5$ indicates perfect alignment between the two distributions.

Figure 1 shows the distribution of $\omega(k)$ across concepts. The first panel serves as a baseline: generating from the same model under identical prompts but different random seeds produces a distribution concentrated around 0.5, confirming that seed variance alone cannot explain the patterns we observe. In most adapted

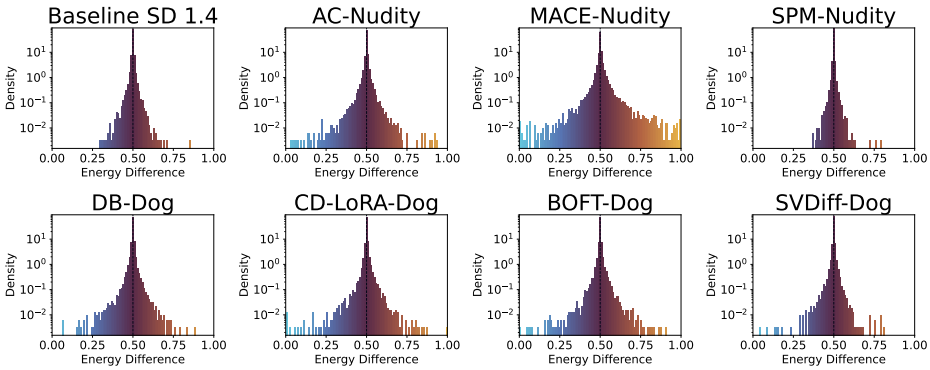


Fig. 1: Distribution of SAE drift scores $\omega(k)$ across concepts for multiple adaptation paradigms. **First panel (top-left):** baseline control (same architecture, same prompts, different random seeds) showing that stochastic sampling variance alone concentrates sharply around 0.5. All remaining panels show adapted variants, where heavy-tailed behavior indicates systematic concept-level drift beyond the adapted target. Top row: concept unlearning (nudity target). Bottom row: concept customization (dog concept). See Appendix A for the full set of distributions across all evaluated methods.

variants, by contrast, we observe heavy-tailed behavior with mass near 0 and 1, indicating systematic concept drift-level drift. This analysis reveals *which* concepts shift, but not whether those shifts translate into a measurable loss of semantic capability. Therefore, we turn to zero-shot classification to answer that question.

Zero-Shot Degradation. If the concept-level shifts surfaced by the SAE analysis are semantically meaningful, they should be reflected in the model's ability to generate recognizable, class-discriminative content. To test this, we evaluate zero-shot classification performance using the Gaussian Diffusion Classifier (GDC) [25]. For each class, GDC generates 240 images and constructs a class prototype by averaging their CLIP image embeddings; test images are then assigned to the nearest prototype by cosine similarity, requiring no real training data at any stage. Classification accuracy therefore depends entirely on the semantic coverage of the generative model: if adaptation erodes a concept, the corresponding prototype degrades and accuracy on that class falls.

We follow the GDC evaluation protocol on three standard benchmarks: CIFAR-10 [15], Oxford Flowers 102 [24], and Food-101 [4]. We report both mean accuracy change and worst-case per-class accuracy drop relative to the base model. The gap between these two statistics is the key diagnostic: a small mean drop alongside a large worst-case drop indicates concentrated semantic damage on specific concepts, invisible to aggregate evaluation. Together, the SAE and GDC analyses establish that adaptation inflicts real, concept-specific damage, providing the empirical foundation for DriftScope, which is designed to surface exactly this kind of hidden drift before deployment.

Table 1: Worst-case per-class accuracy drop and mean accuracy delta induced by model adaptations, measured via zero-shot classification with GDC. We additionally report FID and KID between images generated by the base and modified models under identical prompt distributions. For each method we report the worst single per-class drop and the mean delta across all classes. For unlearning methods, both metrics are computed on a single target concept (nudity). For DreamBooth, we report mean ± std across all fine-tuned concepts.

Area	Method	Model	FID↓	KID↓	CIFAR-10	Flowers-102	Food-101
Unlearning	AC	SD1.4	10.250	0.002	−21.5/−2.05	−87.3/−5.81	−29.5/−2.19
	EraseDiff	SD1.4	321.621	0.377	−91.9/ − 23.24	−98.1/−23.60	−87.3/−5.68
	MACE	SD1.4	60.812	0.033	−39.8/ − 5.93	−98.1/−19.77	−49.1/−11.65
	SPM	SD1.4	3.643	0.00002	−18.9/ − 3.91	−14.6/−0.13	−5.0/−0.19
Customization	DreamBooth	SD1.5	6.86 ± 2.18	0.002 ± 0.001	0.96 ± 7.95	−2.88 ± 12.21	0.08 ± 11.86
	DreamBooth	SD2.1	6.68 ± 3.66	0.003 ± 0.003	−2.22 ± 5.73	−0.74 ± 10.91	−0.91 ± 12.45
	DreamBooth	SD3.5	8.45 ± 2.33	0.003 ± 0.001	0.84 ± 8.88	−0.21 ± 10.72	−1.11 ± 9.36

Table 1 reports zero-shot classification accuracy changes and distribution-level metrics across all evaluated methods. All FID and KID values are computed between images generated by the base and modified models under identical prompt distributions, isolating the effect of adaptation from any pre-existing gap between the model and its training data. For customization methods, zero-shot results are averaged across models fine-tuned on 10 different concepts; the standard deviation across concepts is therefore a direct measure of how unevenly damage is distributed.

The results reveal three distinct regimes of adaptation damage. The first, illustrated by EraseDiff, is catastrophic and detectable: FID and KID are substantially elevated, and zero-shot accuracy collapses across all three benchmarks, with worst-case drops of −91.9 on CIFAR-10 and −98.1 on Flowers. Standard aggregate metrics do surface this failure, but at the cost of rendering the model nearly unusable. The second regime, illustrated by SPM, is more insidious: FID and KID remain low (3.643 and 0.00002, respectively), yet zero-shot accuracy drops of up to −18.9 on CIFAR-10 reveal that specific concepts have been quietly eroded. The model passes aggregate inspection while silently losing semantic coverage. The third regime, exhibited by DreamBooth across all three base architectures, is the most subtle. FID and KID are consistently low, and *mean* accuracy changes are near zero, yet the standard deviation across concepts tells a different story: on Flowers, SD1.5 shows a mean drop of only −2.88 points alongside a standard deviation of 12.21, implying that while most classes are unaffected, some are severely damaged. No aggregate metric can surface this, because the damage is concentrated and concept-specific. It is precisely the failure mode that DriftScope is designed to expose.

3.3 DriftScope: Prompt-Level Model Diffing

Identifying which concepts are most affected by a model adaptation requires a probe that is both semantically grounded and differentiable; one that can navigate the discrete token space of natural language while remaining sensitive to

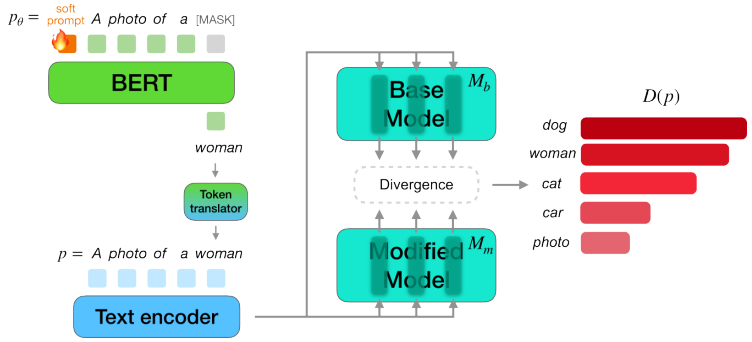


Fig. 2: DriftScope optimizes a soft prompt $p_\theta = [\mathbf{s}_\theta, \mathbf{t}, \mathbf{m}]$ to find the discrete tokens whose visual concepts diverge most between a base model $\mathcal{M}_b$ and its adapted counterpart $\mathcal{M}_m$. The resulting ranked drift report surfaces concepts most affected by fine-tuning.

weight-level changes in a diffusion model. We draw inspiration from two complementary ideas. First, the masked language modeling paradigm [7]: by treating prompt construction as a fill-mask task, we can optimize *which words* to place in a sentence rather than engineering prompts by hand, letting the optimization surface tokens that are maximally informative about a model's behavior. Second, differentiable prompt optimization methods [6,19], which have shown that mapping learnable continuous embeddings to discrete tokens via Gumbel-Softmax relaxation can effectively probe the internal representations of vision and language models. These methods optimize prompts to explain or steer a *single* model; we repurpose the same paradigm for an entirely different goal: *comparing two model checkpoints* to localize where adaptation has caused semantic drift.

DriftScope takes a base diffusion model $\mathcal{M}_b$ and its adapted counterpart $\mathcal{M}_m$ and learns a soft prompt whose discrete realization maximizes the divergence between the cross-attention maps produced by each checkpoint under identical noise seeds, depicted in Figure 2. Rather than asking "what does this model focus on?", we ask "where do these two models disagree?". Once the prompt is optimized, divergence is attributed token-by-token, yielding a ranked blacklist of words whose associated visual concepts have shifted most between checkpoints. Because the method is prompt-driven and operates on any pair of checkpoints, it surfaces non-obvious collateral damage (concepts the practitioner never intended to modify) providing an interpretable, concept-level audit before deployment.

Prompt Parameterization and Encoder Translation. We parameterize the input to the masked language model as a sequence $p_\theta = [\mathbf{s}_\theta, \mathbf{t}, \mathbf{m}]$, where $\mathbf{s}_\theta = \{s_i\}$ is a set of learnable soft prompt vectors prepended to the sequence, $\mathbf{t} = \{t_j\}$ is the tokenized fixed text, and $\mathbf{m} = \{m_k\}$ are the mask tokens whose discrete realizations we wish to optimize. The masked language model predicts a distribution over vocabulary entries for each m_k; we obtain a differentiable dis-

crete token via Gumbel-Softmax relaxation, enabling gradient-based optimization of $\mathbf{s}$ through standard backpropagation.

Because diffusion architectures employ different text encoders (*e.g.*, CLIP variants in SD 1/2/3/XL, T5 in SD3 and FLUX), the tokens predicted in the masked language model's vocabulary may not map directly into the target encoder's vocabulary. To bridge this gap, we employ a token translator: a fixed binary matrix $M \in \{0,1\}^{V \times W}$ that maps each one-hot token vector from the source vocabulary of size V to its closest counterpart in the target vocabulary of size W. This design, used in prior prompt optimization work [6, 27], allows DriftScope to operate across diffusion architectures without modifying the divergence objective or the optimization procedure.

During optimization, both diffusion model checkpoints remain frozen; gradients flow only through the soft prompt parameters $\mathbf{s}_\theta$. To ensure that the tokens discovered generalize beyond a single stochastic realization, we optimize over a batch of S fixed noise seeds rather than a single sample, so that high-scoring tokens reflect consistent cross-model divergence rather than the idiosyncrasies of a particular noise realization.

DriftScope Objective. We aim to find a prompt that maximally exposes the representational differences induced by adaptation, by measuring how differently the two checkpoints process the same textual input across noise seeds.

Cross-attention extraction. For a prompt p and noise seed z, both models denoise latent trajectories under the same sampling schedule. At each step, the network produces cross-attention maps that relate textual tokens to spatial latent features. Depending on the architecture (UNet-based or transformer-based) layers may operate at varying or identical spatial resolutions. We denote the cross-attention map for a given layer as:

$$A(\mathcal{M}, p, z) \in \mathbb{R}^{H \times W \times T}, \tag{2}$$

where $H \times W$ is the spatial resolution and T the number of prompt tokens. The subset of layers $\mathcal{L}$ most sensitive to adaptation drift is architecture-dependent; for SD1.5, for instance, this corresponds to `down_block_0` through `up_block_3`.

Divergence objective. We define the divergence for a given prompt as the L_1 difference between cross-attention maps, averaged across the seed batch $\{z_1, \ldots, z_S\}$ and layers $\mathcal{L}$:

$$D(p) = \frac{1}{S} \sum_{s=1}^{S} \sum_{\ell \in \mathcal{L}} \left\| A^{(\ell)}(\mathcal{M}_b, p, z_s) - A^{(\ell)}(\mathcal{M}_m, p, z_s) \right\|_1 . \tag{3}$$

The L_1 norm was chosen over L_2 as it produced sharper token-level localization and greater stability across seeds.

Prompt optimization. We optimize the soft prompt $\mathbf{s}_\theta$ to maximize D:

$$\mathbf{s}_\theta^* = \arg\max_{\mathbf{s}_\theta} D(p_\theta), \tag{4}$$

where $p_\theta = [\mathbf{s}_\theta, \mathbf{t}, \mathbf{m}]$ follows the parameterization introduced in the prompt parametrization definition. Gradients are backpropagated through both frozen checkpoints simultaneously.

Word-Level Drift Attribution. Once the prompt $p_\theta^* = [\mathbf{s}_\theta^*, \mathbf{t}, \mathbf{m}^*]$ has been optimized, DriftScope assigns a drift score to the discrete tokens realized at the mask positions $\mathbf{m}^*$. Rather than isolating each token's attention slice, the score is computed over the cross-attention maps of the full prompt, capturing how the two checkpoints process the entire scene (including object relations and context) rather than individual concept appearances:

$$\Delta(\mathbf{m}^*) = \frac{1}{S}\sum_{s=1}^{S}\sum_{\ell\in\mathcal{L}}\left\| A^{(\ell)}(\mathcal{M}_b, p^*, z_s) - A^{(\ell)}(\mathcal{M}_m, p^*, z_s)\right\|_1. \tag{5}$$

By running the optimization multiple times with different initializations, each run yields a candidate token (or set of tokens) together with its associated drift score Δ. Because scores are aggregated across multiple seeds and layers, the ranking reflects consistent representational change rather than noise from any particular stochastic realization.

Practical Usage. In practice, a user provides a prompt of interest and two checkpoints; DriftScope returns a ranked list of tokens whose visual concepts exhibit the largest cross-model divergence, or effectively, a blacklist of concepts that have been most affected by adaptation. By inverting the objective, DriftScope can equally identify tokens that minimize Δ, revealing concepts that remain stable across checkpoints. For deployment scenarios, practitioners can run DriftScope over a set of anticipated prompts and aggregate the most frequently appearing high-drift tokens, surfacing systematically unstable concepts before an adapted model is released.

4 Experiments

4.1 Experimental setup

Evaluation Protocol. We evaluate DriftScope by measuring whether the tokens it identifies as high-drift produce maximally divergent image distributions between the base and modified models.

Template Prompts. We choose five template prompts from Textual Inversion [9] formulations (*e.g.*, "A photo of a `[MASK]`"). These templates provide controlled semantic contexts while allowing DriftScope to optimize the target concept token (see Supplementary Material for details).

Drift Optimization. For each template, we run the DriftScope optimization procedure 100 times with different prompt initialization seeds, yielding 500 optimization runs per method. To reduce memory requirements and wall-clock time, we restrict denoising to the first 4 steps of the diffusion trajectory; early steps govern high-level semantic structure and are most sensitive to concept-level drift [2], making further steps unnecessary for reliable drift signal. In each run, we extract the top-ranked drift tokens according to the word-level divergence score defined in Section 3.3. We then aggregate results across runs and select the five most frequently occurring high-drift tokens per method.

Image Generation. For each selected token, we generate 100 images from each checkpoint under identical sampling conditions (seeds, resolution, guidance scale, and number of denoising steps), ensuring that observed differences arise solely from the checkpoint weights. DreamBooth models are fine-tuned on the publicly available dog images provided by the DreamBooth authors (shown in Supplementary Material). Unlearning checkpoints are taken from the AdvUnlearn repository*.

Distribution Comparison. This procedure produces paired image sets for each high-drift token. We quantify divergence between the base and modified image distributions using the metrics described below. All experiments are carried on identical prompt distributions and sampling configurations across methods.

Metrics. We quantify divergence between image sets generated by the base and modified models using four complementary metrics. CLIP-I measures the average cosine similarity between CLIP image embeddings of corresponding pairs generated under identical prompts and noise seeds — lower values indicate greater semantic divergence. MS-SWD [12] captures discrepancies in multi-scale image statistics, including structural and textural differences, with higher values indicating greater distributional divergence. LPIPS [36] measures perceptual distance using deep feature activations from a pretrained network, correlating more closely with human judgments than pixel-wise metrics, with higher values indicating greater perceptual divergence. Finally, Q-Eval [37] evaluates prompt-image alignment and perceptual quality using a unified metric trained on large-scale human preference annotations; a drop in score indicates that adaptation has degraded the model's ability to generate a concept both perceptually and semantically.

4.2 Quantitative Results

We evaluate DriftScope on a representative subset spanning different damage regimes: ESD [10] and SPM [22] for concept unlearning, Scissorhands (SH) [33] as a degenerate-regime baseline, and DreamBooth across three base architectures

* `https://github.com/OPTML-Group/AdvUnlearn?tab=readme-ov-file#checkpoints`

for concept customization. For each method, we evaluate across 6 concepts spanning diverse semantic categories. DriftScope identifies the five most frequently occurring high-drift tokens across 500 optimization runs per concept.

Tables 2 and 3 jointly structure the evaluation: Table 2 reports metrics for high-drift tokens surfaced by DriftScope, while Table 3 serves as an internal reference for low-drift tokens identified by inverting the objective. If DriftScope is working, the gap between the two tables should be consistent and interpretable. The ranked token lists for each method and concept are provided in Appendix D.

The clearest illustration comes from SPM. High-drift tokens yield a CLIP-i of 0.88 and LPIPS of 0.33, while low-drift tokens reach 0.96 and 0.14 respectively, a sharp separation that confirms DriftScope is reliably discriminating between stable and unstable concepts. The words themselves reinforce the paper's central claim, as unlearning nudity surfaces *body, goddess, woman* as collaterally damaged concepts that the practitioner never intended to modify**.

ESD presents a more complex picture. The high-drift vocabulary (*seal, cross, boyfriend*) is semantically distant from the nudity target, illustrating non-obvious collateral damage, yet the gap between high- and low-drift conditions is modest across metrics, suggesting ESD induces broader, less localized damage that makes clean separation harder.

Scissorhands represents the degenerate regime described in Section 3: a Q-Eval of 0.12 and CLIP-i of 0.56 on high-drift tokens indicate widespread model degradation rather than targeted concept erosion. As discussed in the Limitations, DriftScope's attributions for this method should be interpreted with caution.

For DreamBooth, the most informative axis is Q-Eval, which captures both perceptual quality and prompt faithfulness. As shown in Table 1, mean accuracy changes are near zero across all three architectures, yet the standard deviation reveals concept-specific instability. Accordingly, SD2.1 drops to 0.39 on high-drift tokens while SD3.5 remains at 0.69, suggesting the latter is more robust to collateral damage. The drift vocabulary is also telling: SD3.5 surfaces *bible, dragon, devil* as the concepts most affected by a dog fine-tune, while SD2.1 includes *dog* itself alongside *woman* and *car*, all examples of non-obvious collateral damage on concepts the practitioner never intended to modify.

A key property of DriftScope is that the tokens it surfaces should reflect the specific concept being adapted rather than a generic fine-tuning signature. The results confirm that collateral damage is structured and target-dependent: the drift vocabulary shifts predictably with the adapted concept, tracing damage back to its source in the model's representational space. Crucially, this does not mean the damaged concepts are obvious or intended. Unlearning *garbage* displaces *truck* and *bus*, while unlearning *tench* affects *book* and *calendar*, none of which a practitioner would anticipate. This target-dependence is precisely what makes DriftScope actionable: it does not return a generic blacklist, but a drift report that is specific to each adaptation.

** The high-drift tokens can be found in the Supplementary material Table 4

Table 2: High-drift tokens and image-level divergence metrics identified by DriftScope.

Method	Model	CLIP-i ↑	MS-SWD↓	LPIPS↓	Q-Eval↑
Unlearning					
ESD	SD1.4	0.72 ± 0.13	1.84	0.57 ± 0.15	0.41 ± 0.17
SPM	SD1.4	0.88 ± 0.11	1.02	0.33 ± 0.20	0.50 ± 0.16
SH	SD1.4	0.56 ± 0.09	4.15	0.82 ± 0.08	0.12 ± 0.07
Customization					
Dreambooth	SD1.5	0.81 ± 0.13	1.87	0.51 ± 0.18	0.52 ± 0.18
Dreambooth	SD2.1	0.81 ± 0.10	1.81	0.54 ± 0.14	0.39 ± 0.15
Dreambooth	SD3.5	0.81 ± 0.10	2.20	0.54 ± 0.13	0.69 ± 0.17

Table 3: Low-drift tokens and image-level divergence metrics identified by DriftScope.

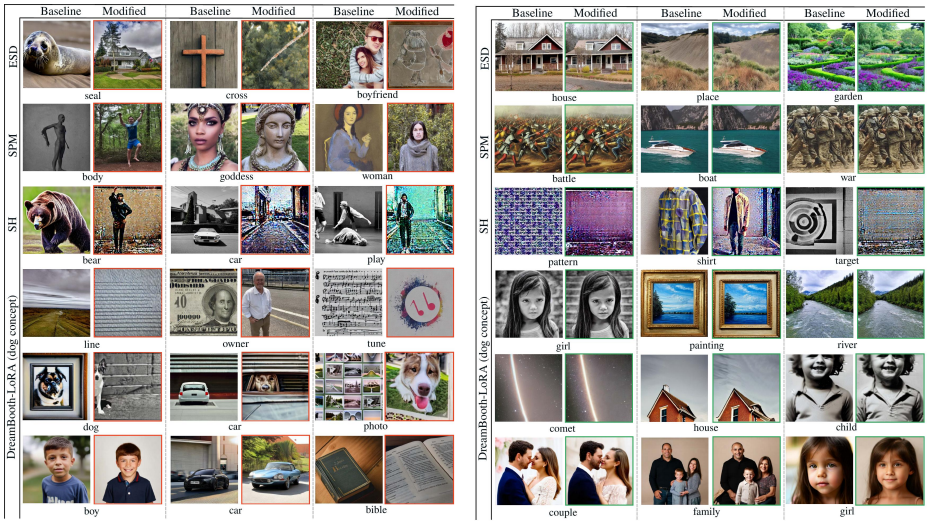
Method	Model	CLIP-i ↑	MS-SWD↓	LPIPS↓	Q-Eval↑
Unlearning					
ESD	SD1.4	0.79 ± 0.11	1.74	0.52 ± 0.13	0.41 ± 0.12
SPM	SD1.4	0.96 ± 0.05	0.39	0.14 ± 0.11	0.43 ± 0.14
SH	SD1.4	0.62 ± 0.10	4.01	0.78 ± 0.11	0.19 ± 0.08
Customization					
Dreambooth	SD1.5	0.85 ± 0.10	1.41	0.44 ± 0.16	0.56 ± 0.16
Dreambooth	SD2.1	0.84 ± 0.09	1.44	0.50 ± 0.15	0.42 ± 0.16
Dreambooth	SD3.5	0.81 ± 0.10	2.11	0.53 ± 0.13	0.71 ± 0.16

4.3 Qualitative Results

Figure 3 provides the visual counterpart to the quantitative results. High-drift pairs show changes well beyond stylistic variation: for ESD, *seal* shifts from a marine animal to a house exterior and *boyfriend* transforms from a portrait to a sketch; for SPM, *body* and *goddess* both shift markedly in composition and style, suggesting nudity unlearning broadly reshapes human-form rendering. In DreamBooth, fine-tuning bleeds into adjacent tokens such as *bible* and *car* in SD3.5 and *dog* itself in SD2.1. Low-drift tokens (SPM's *battle*, *boat*; DreamBooth's *girl*, *painting*) produce near-identical pairs, confirming DriftScope discriminates between stable and unstable concepts. Scissorhands outputs remain degraded in both conditions, consistent with the widespread model collapse identified quantitatively.

5 Ablations

Generalization to Long and Domain-Specific Prompts. DriftScope is not limited to short, single-concept prompts. Figure 4 shows that collateral damage surfaces equally in longer, naturalistic inputs: in the DreamBooth dog fine-tune, *Facebook* emerges as a high-drift token despite having no relation to the adapted concept,



(a) Maximized drift. **(b) Minimized drift.**

Fig. 3: Qualitative examples of prompts found by DriftScope that (a) *maximize* and (b) *minimize* cross-attention drift between model checkpoints. We report results for unlearning methods (ESD, SPM, and Scissorhands) and DreamBooth LoRA customization (dog concept) across Stable Diffusion 1.5, 2.1, and 3.5-Medium.

and color-bearing prompts show stylistic drift propagating to modifiers the practitioner would never expect to be affected. This confirms that DriftScope's utility extends to the open-ended prompts encountered in real deployment.

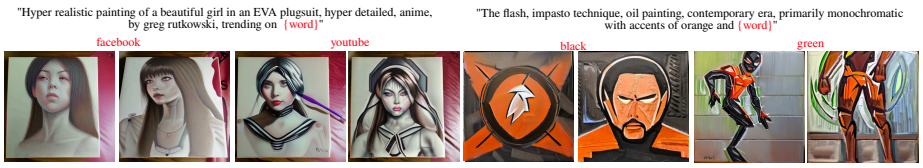


Fig. 4: DriftScope generalizes beyond short single-concept prompts. Shown are high-drift image pairs from SD 2.1 under long, naturalistic inputs sourced from PromptHero [1], for a DreamBooth model fine-tuned on a dog concept. Each pair shows the base model generation (left) alongside the adapted model generation (right).

6 Limitations

DriftScope is optimization-based rather than analytic, searching across many initializations and averaging over multiple seeds, which makes it computationally intensive while mitigating sensitivity to degenerate noise realizations. The method also assumes both checkpoints retain meaningful generative structure:

when adaptation causes widespread degradation (as with Scissorhands), cross-attention maps become globally uninformative and concept-level attribution breaks down.

7 Conclusion

Adapting text-to-image diffusion models is routinely treated as a benign operation, evaluated only on its intended effects. We have shown this framing is insufficient. Across concept customization and concept unlearning weight-level modification systematically damages semantically unrelated concepts in ways that aggregate metrics structurally cannot surface. When the damage is severe enough for FID and KID to respond, the model is already nearly unusable; when the model remains functional, FID and KID stay flat while specific classes silently suffer worst-case zero-shot accuracy drops of up to 18.9 points and concept-level distributions shift dramatically.

Building on this evidence, we introduced DriftScope, a prompt-level diagnostic that identifies which words have their visual concepts most affected by adaptation. By maximizing divergence between the cross-attention maps of two checkpoints, DriftScope surfaces non-obvious collateral damage and returns a ranked blacklist of high-drift tokens for any user-provided prompt, without requiring access to real data or model internals beyond standard attention outputs.

Acknowledgments

This work was supported by Grants PID2022-143257NB-I00, PID2024-162555OB-I00, and AIA2025-163919-C52 funded by MICIU/AEI/10.13039/501100011033, ERDF/EU and the FEDER, by the Generalitat de Catalunya CERCA Program, by the grant Càtedra ENIA UAB-Cruïlla (TSI-100929-2023- 2) from the Ministry of Economic Affairs and Digital Transition of Spain, by the European Union's Horizon Europe research and innovation programme under grant agreement number 101214398 (ELLIOT), and by the BBVA Foundations of Science program on Mathematics, Statistics, Computational Sciences and Artificial Intelligence (grant VIS4NN). Kai Wang acknowledges the funding No.R002026G0116 and No.R002026B0121 from Guangdong and Hong Kong Universities 1+1+1 Joint Research Collaboration Scheme, and the start-up grant B01040000108 from CityU-DG. Yiping Han acknowledges the Chinese Scholarship Council (CSC) grant No.202504910071. We acknowledge the EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, and the RES resources provided by BSC on MareNostrum5 under projects IM-2025-3-0025 and IM-2025-3-0027.

References

1. PromptHero. https://prompthero.com, accessed: 2026
2. Biroli, G., Bonnaire, T., de Bortoli, V., Mézard, M.: Dynamical regimes of diffusion models. Nature Communications (2024)
3. Bohacek, M., Fel, T., Agrawala, M., Lubana, E.S.: Uncovering conceptual blindspots in generative image models using sparse autoencoders. arXiv preprint arXiv:2506.19708 (2025)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision. pp. 446–461. Springer (2014)
5. Bricken, T., et al.: Stage-wise model diffing. https://transformer-circuits.pub/2024/model-diffing/index.html (2024), anthropic Interpretability Team
6. Carnemolla, S., Pennisi, M., Samarasinghe, S., Bellitto, G., Palazzo, S., Giordano, D., Shah, M., Spampinato, C.: Dexter: Diffusion-guided explanations with textual reasoning for vision models. Advances in neural information processing systems (2025)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dunlap, L., Gonzalez, J.E., Darrell, T., Heilbron, F.C., Sivic, J., Russell, B.: Discovering divergent representations between text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17516–17525 (2025)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: International Conference on Learning Representations (ICLR) (2023)
10. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2426–2436 (2023)
11. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7323–7334 (2023)
12. He, J., Wang, Z., Wang, L., Liu, T.I., Fang, Y., Sun, Q., Ma, K.: Multiscale sliced Wasserstein distances as perceptual color difference measures. In: European Conference on Computer Vision. pp. 1–18 (2024), http://arxiv.org/abs/2407.10181
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Kaushik, A.R., Devulapally, N.K., Lokhande, V.S., Ratha, N., Govindaraju, V.: Forget less by learning together through concept consolidation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 265–275 (2026)
15. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
16. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. CVPR (2023)
17. Laria, H., Gomez-Villa, A., Wang, K., Raducanu, B., van de Weijer, J.: Assessing open-world forgetting in generative image model customization. arXiv preprint arXiv:2410.14159 (2024)

18. Lindsey, J., et al.: Crosscoders: A sparse autoencoder architecture for comparing model representations. https://transformer-circuits.pub/2024/crosscoders/index.html (2025), anthropic Interpretability Team
19. Liu, Q., Kortylewski, A., Bai, Y., Bai, S., Yuille, A.: Discovering failure modes of text-guided diffusion models via adversarial search. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=TOWdQQgMJY
20. Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., Feng, H., Liu, Z., Heo, J., Peng, S., Wen, Y., Black, M.J., Weller, A., Schölkopf, B.: Parameter-efficient orthogonal finetuning via butterfly factorization. In: ICLR (2024)
21. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
22. Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., Ding, G.: One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
23. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., Van De Weijer, J.: Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5513–5533 (2022)
24. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
25. Qi, Z., Liu, B., Zhang, S., Li, B., Xu, Z., Xiong, H., Xie, Z.: A simple and efficient baseline for zero-shot generative classification. arXiv preprint arXiv:2412.12594 (2024)
26. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
27. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 4222–4235 (2020)
28. Smith, J.S., Hsu, Y.C., Zhang, L., Hua, T., Kira, Z., Shen, Y., Jin, H.: Continual diffusion: Continual customization of text-to-image diffusion with c-lora. Transactions on Machine Learning Research (2024)
29. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265. PMLR (2015)
30. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
31. Tong, S., Jones, E., Steinhardt, J.: Mass-producing failures of multimodal systems with language models. Advances in neural information processing systems **36**, 29292–29322 (2023)
32. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=lm7MRcsFiS

33. Wu, J., Harandi, M.: Scissorhands: Scrub data influence via connection sensitivity in networks. In: European Conference on Computer Vision. pp. 367–384. Springer (2024)
34. Wu, J., Le, T., Hayat, M., Harandi, M.: Erasing undesirable influence in diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28263–28273 (2025)
35. Zajac, M., Deja, K., Kuzina, A., Tomczak, J.M., Trzcinski, T., Shkurti, F., Milo's, P.: Exploring continual learning of diffusion models. ArXiv **abs/2303.15342** (2023), https://api.semanticscholar.org/CorpusID:257766772
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
37. Zhang, Z., Kou, T., Wang, S., Li, C., Sun, W., Wang, W., Li, X., Wang, Z., Cao, X., Min, X., et al.: Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10621–10631 (2025)

A Full SAE Drift Score Distributions

Figure 5 extends Figure 1 to all evaluated adaptation paradigms and methods. The baseline panel (top-left) is reproduced for reference. All remaining panels show adapted variants grouped by paradigm: concept unlearning (nudity target) and concept customization (dog concept). In all cases the baseline concentrates sharply around $\omega(k) = 0.5$, while adapted variants exhibit heavy-tailed behavior to varying degrees, confirming that the phenomenon is not specific to any single method or paradigm.

B DreamBooth Fine-Tuning Data

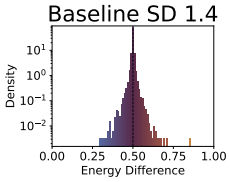
Figure 6 shows the five images used to fine-tune the DreamBooth models evaluated in this paper, taken from the publicly available dog dataset provided by the DreamBooth authors [26].

C Template Prompts

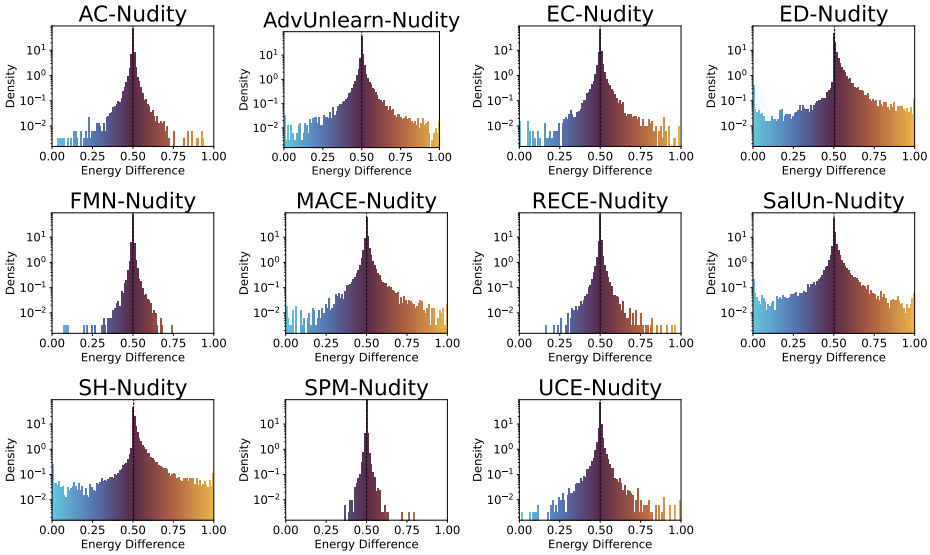
Below we provide the template prompts chosen from Textual Inversion [9]:

- A photo of a [MASK]
- A rendition of the [MASK]
- A photo of a dirty [MASK]
- A rendering of a [MASK]
- A dark photo of a [MASK]

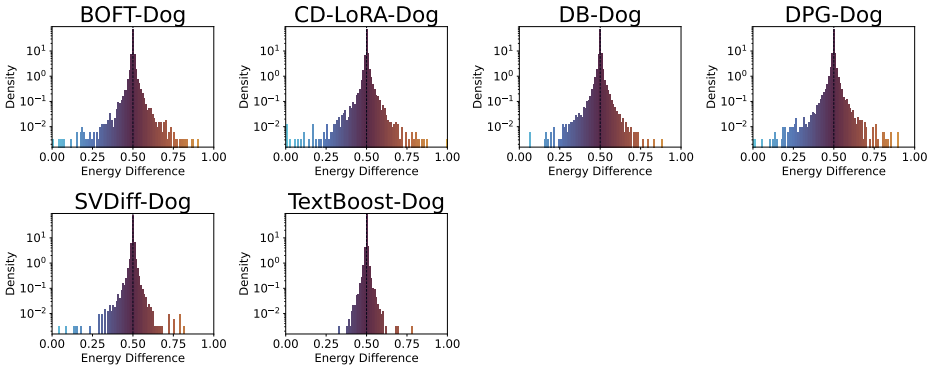
D Top words for each concept



Baseline (seed variance control)



Concept unlearning (nudity target): AC, AdvUnlearn, EC, EraseDiff, FMN, MACE, RECE, SalUN, Scissorhands, SPM, UCE



Concept customization (dog concept): BOFT, CD-LoRA, DreamBooth, DPG, SVDiff, TextBoost

Fig. 5: Full distribution of SAE drift scores $\omega(k)$ across all evaluated methods and paradigms. The baseline panel (top-left) is reproduced from Figure 1 for reference. Within each paradigm, panels are ordered alphabetically by method name. Heavy-tailed behavior near 0 and 1 indicates systematic concept-level drift; the degree and shape of the tails vary across methods but the pattern is consistent within each paradigm.

Fig. 6: The five dog images used for DreamBooth fine-tuning, sourced from the publicly available dataset released by [26].

Table 4: High-drift tokens identified by DriftScope on models adapted to unlearn different concepts. Each row corresponds to an independently adapted checkpoint.

Method	Model	Concept	Top words
ESD	SD1.4	Church	wolf, russian, reporter, future, beach
		Garbage	blond, blonde, baby, bag, tattoo
		Parachute	child, text, conversation, words, daughter
		Tench	blond, blonde, russian, artist, professor
		Van Gogh	child, text, words, friend, pair
		Nudity	target, cross, seal, boyfriend, pattern
SPM	SD1.4	Church	church, sermon, house, funeral, priest
		Garbage	truck, opera, laundry, bus, musical
		Parachute	wind, angel, fairy, comet, balloon
		Tench	sun, fish, tree, prince, poet
		Van Gogh	painting, painter, poet, picasso, moon
		Nudity	body, woman, baby, goddess, sheet
SH	SD1.4	Church	map, meeting, room, tray, wall
		Garbage	table, room, cast, classroom, meeting
		Parachute	band, map, meeting, show, table
		Tench	book, calendar, map, meeting, room
		Van Gogh	crowd, painting, target, vamp, wall
		Nudity	car, camera, horse, play, bear
DreamBooth	SD1.5	furniture_sofa	target, poster, dozen, bible, baby
		jewelry_ring	knife, target, notebook, anthem, laptop
		plushie_panda	text, cat, baby, guy, words
		things_book	target, note, page, text, book
		toy_unicorn	bible, monster, demon, David, target
		pets_dog	baby, tune, owner, body, line
DreamBooth	SD2.1	furniture_sofa	man, face, car, portrait, song
		jewelry_ring	song, flag, towel, book, words
		plushie_panda	picture, dozen, man, voice, guy
		things_book	picture, melody, target, photograph, name
		toy_unicorn	song, picture, target, name, tune
		pets_dog	dog, woman, cat, car, photo
DreamBooth	SD3.5	decoritems_lamp	man, girl, woman, cat, family
		person	woman, man, girl, picture, party
		scene_lighthouse	girl, man, face, couple, vampire
		scene_waterfall	girl, couple, vampire, woman, poem
		transport_car	girl, woman, man, child, boy
		pets_dog	bible, dragon, car, report, devil

Table 5: Low-drift tokens identified by DriftScope on models adapted to unlearn different concepts. Each row corresponds to an independently adapted checkpoint.

Method	Model	Concept	Top words
ESD	SD1.4	Church	boy, shirt, window, harp, melody
		Garbage	house, tree, mansion, pond, mountain
		Parachute	city, car, holocaust, church, movie
		Tench	house, car, tree, knife, beatles
		Van Gogh	house, funeral, piano, choir, mass
		Nudity	garden, house, bitch, place, funeral
SPM	SD1.4	Church	tune, figure, duck, horse, worker
		Garbage	picture, horse, spider, book, photo
		Parachute	photo, picture, man, photograph, car
		Tench	car, cop, hotel, plane, band
		Van Gogh	car, shop, title, bible, package
		Nudity	car, war, dog, boat, battle
SH	SD1.4	Church	woman, devil, stranger, girl, cross
		Garbage	devil, woman, cross, suicide, female
		Parachute	female, cross, stranger, soldier, patient
		Tench	devil, stranger, cross, woman, female
		Van Gogh	man, nun, boy, biography, brother
		Nudity	target, shirt, text, bible, pattern
DreamBooth	SD1.5	furniture_sofa	cat, doctor, nurse, baby, train
		jewelry_ring	ship, couple, nurse, man, train
		plushie_panda	restaurant, camera, statue, photo, room
		things_book	house, stranger, shop, woman, life
		toy_unicorn	room, spider, building, night, wall
		pets_dog	landscape, painting, scene, river, girl
DreamBooth	SD2.1	furniture_sofa	tempest, baby, spider, bed, flower
		jewelry_ring	opera, tempest, cat, machine, kitchen
		plushie_panda	scene, novel, bed, stranger, baby
		things_book	painting, party, boy, sermon, mother
		toy_unicorn	bible, scene, tomb, body, bathroom
		pets_dog	face, couple, house, child, comet
DreamBooth	SD3.5	decoritems_lamp	child, woman, girl, vampire, boy
		person	girl, woman, car, kid, novel
		scene_lighthouse	woman, girl, man, boy, face
		scene_waterfall	girl, vampire, woman, child, face
		transport_car	girl, man, boy, baby, woman
		pets_dog	couple, man, boy, girl, family