

# Distill, Forget, Repeat: A Framework for Continual Unlearning in Text-to-Image Diffusion Models

Naveen George<sup>1,2,†</sup>, Naoki Murata<sup>2</sup>, Yuhta Takida<sup>2</sup>, Konda Reddy Mopuri<sup>1</sup>, Yuki Mitsufuji<sup>2,3</sup>

<sup>1</sup>Indian Institute of Technology Hyderabad <sup>2</sup>Sony AI <sup>3</sup>Sony Group Corporation

ai23mtech12001@iith.ac.in, naoki.murata@sony.com

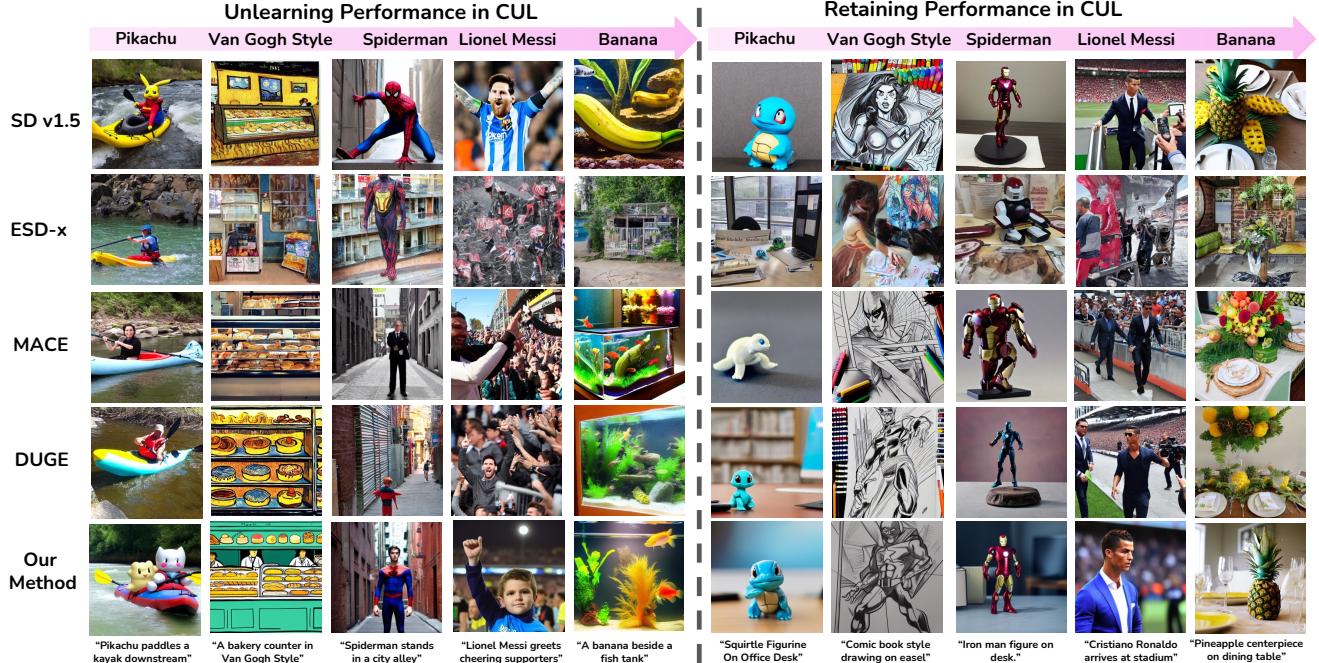


Figure 1. Qualitative comparison of our method (bottom row) against SOTA baselines on 10 sequential unlearning steps. This figure highlights the critical failure modes of existing methods in a continual setting. Baselines like ESD-x [9] and MACE [19] suffer from catastrophic “retention collapse”, where generative quality completely breaks down on retained concepts (right panel). Other methods like DUGE [32] avoid total collapse but still exhibit severe quality degradation and poor unlearning in later stages. In contrast, our method demonstrates a superior ability to both effectively unlearn target concepts (left panel) and preserve general knowledge (right panel), maintaining a good generative quality compared to the original SD model.

## Abstract

The recent rapid growth of visual generative models trained on vast web-scale datasets has created significant tension with data privacy regulations and copyright laws, such as GDPR’s “Right to be Forgotten.” This necessitates machine unlearning (MU) to remove specific concepts without the prohibitive cost of retraining. However, existing MU techniques are fundamentally ill-equipped for real-world scenarios where deletion requests arrive sequentially, a setting known as continual unlearning (CUL). Naively applying one-shot methods in a continual setting triggers a stability

crisis, leading to a cascade of degradation characterized by retention collapse, compounding collateral damage to related concepts, and a sharp decline in generative quality. To address this critical challenge, we introduce a novel generative distillation based continual unlearning framework that ensures targeted and stable unlearning under sequences of deletion requests. By reframing each unlearning step as a multi-objective, teacher-student distillation process, the framework leverages principles from continual learning to maintain model integrity. Experiments on a 10-step sequential benchmark demonstrate that our method unlearns forget concepts with better fidelity and achieves this without significant interference to the performance on retain concepts or the overall image quality, substantially outperform-

<sup>†</sup>Work done during an internship at Sony AI

ing baselines. This framework provides a viable pathway for the responsible deployment and maintenance of large-scale generative models, enabling industries to comply with ongoing data removal requests in a practical and effective manner.

## 1. Introduction

Recent visual generative models such as SORA [5], Gemini [31], Imagen 3.0 [28], and DALL-E 3 [3] can synthesize photorealistic media at scale. These models are trained on massive web-scale datasets that often include sensitive personal data, NSFW material, and copyrighted content. Laws such as GDPR [24] grant a “**Right to be Forgotten**”, but retraining foundation models from scratch is prohibitively expensive. The common workaround, post-generation filtering, is superficial; the model retains underlying knowledge, and filters can be bypassed by users [23, 36]. This motivates machine unlearning (MU): directly editing model weights to remove targeted concepts (objects, NSFW, artistic styles, identities) from generative behavior. However, deletion requests arrive sequentially rather than all at once.

In principle, an ideal unlearning method should satisfy three desiderata: **(I) Perfect unlearning**, where forget concepts are not produced; **(II) No Ripple Effects** [1], so that related concepts remain intact (e.g., removing Brad Pitt should not distort Angelina Jolie or Leonardo DiCaprio; unlearning Van Gogh should not degrade other artistic styles); and **(III) Quality Preservation**, maintaining overall image quality. Current approaches map forget concepts to fixed placeholders (empty strings [9, 34], general concepts [10, 18], or random concepts [8]), suppressing rather than truly unlearning them. This leads to ripple effects (collateral damage to related concepts [1, 40]).

These challenges are significantly amplified in **continual unlearning (CUL)** [11, 32, 35], where deletion requests arrive sequentially over time. When existing one-shot methods are applied repeatedly, they trigger three catastrophic failures that compound across steps: **(1) Retention collapse**, where weak preservation mechanisms (simple KL divergence or anchor mapping [10, 13, 18]) cause cumulative knowledge loss, eventually leading to complete loss of general capabilities [32, 40]; **(2) Compounding ripple effects**, where broad parameter updates [9, 34] and fixed placeholder mapping cause collateral damage that accumulates with each step, blurring semantic boundaries; and **(3) Cumulative parameter drift**, where parameters progressively drift from their original state without explicit regularization [17], destabilizing the model and causing quality degradation. We formalize these as three failure modes in Section 3.2.3. These interacting failures also give rise to concept revival, where previously unlearned concepts spontaneously reappear after later updates [12]. These problems

fundamentally compound in CUL, as each step inherits and magnifies previous errors.

To address these challenges, we propose a distillation-based framework grounded in continual learning principles. We formulate each unlearning step as teacher-student distillation, where the previous-step model serves as teacher providing reference for retained knowledge. Our framework consists of three synergistic components: **(1) Generative replay with distillation** counters retention collapse by explicitly distilling retain concepts using teacher-generated synthetic data, providing stronger knowledge preservation than simple KL constraints; **(2) Parameter regularization** counters drift through L2 penalties constraining deviation from the previous step, achieving stabilization similar to Elastic Weight Consolidation [17]; and **(3) Contextual trajectory re-steering** achieves effective unlearning while mitigating ripple effects through context-preserving mapping strategies instead of fixed placeholders, surgically modifying generation trajectories to minimize collateral damage.

We validate our approach on a challenging 10-step sequential unlearning benchmark. As shown in Figure 1, while existing methods collapse after 3-5 steps with catastrophic quality degradation and concept revival, our method remains stable after 10 sequential steps, successfully unlearning forget concepts (Pikachu, Brad Pitt, Dog, Golf Ball, Van Gogh Style, Apple, Spiderman, Lionel Messi, Cartoon Style, Banana) while retaining general knowledge and image quality. We demonstrate clear superiority over state-of-the-art baselines across all evaluation metrics.

Our contributions are:

- 1. Systematic analysis of CUL challenges:** We elucidate how existing one-shot unlearning methods fail when applied sequentially, identifying retention collapse, ripple effects, and parameter drift as fundamental challenges.
- 2. Distillation-based framework:** We propose a principled framework with three complementary components: generative replay, parameter regularization, and contextual trajectory re-steering, each addressing specific failure modes.
- 3. Evaluation protocol for CUL:** We establish refined metrics and benchmark construction for evaluating continual unlearning.
- 4. Practical solution:** We demonstrate a practical approach enabling industry deployment to handle sequential deletion requests while maintaining model stability and quality.

## 2. Related Works

**Unlearning in Text-to-Image Diffusion Models** Machine unlearning (MU) has emerged as a critical post-hoc technique for aligning large-scale generative models with evolving safety, privacy, and copyright standards, obviating the

need for complete model retraining. Initial approaches focused on re-mapping “forget” concepts to neutral anchors, as pioneered by ESD [9] using classifier-free guidance and formalized by Concept Ablation [18] using KL divergence. However, these holistic modifications often require extensive parameter updates, causing a “ripple effect” [1, 40], or collateral damage, on related concepts. This problem drove the field toward more surgical, parameter-efficient interventions that target small, critical parameter subsets. For instance, some methods directly target the cross-attention layers, such as UCE’s [10] closed-form solution and FMN’s [37] attention re-steering loss.

Another prominent direction, inspired by PEFT, trains lightweight modules while freezing the base model, such as SPM’s [20] “semi-permeable” adapter, Receler’s [16] “eraser” module, or MACE’s [19] use of LoRA. Other surgical approaches identify critical parameters via saliency analysis (SalUn [8]) or by targeting specific denoising timesteps (SHS [33]), while newer methods focus on highly constrained updates by handling “forget” and “retain” gradient conflicts [4, 22, 28]. Despite this progress, these techniques remain fragile, as concepts are often just suppressed. Models remain vulnerable to adversarial attacks [15, 39], persist in the latent space [27], and some work has begun exploring robustness [30, 38].

**Knowledge Distillation** Knowledge distillation (KD) is a technique for transferring knowledge from a large “teacher” model to a smaller “student,” proving to be a powerful tool for knowledge preservation [14]. In diffusion models, existing studies use knowledge distillation to train a student network that can generate the same quality images as a teacher network, but in fewer generation steps. This paradigm has been adapted for the continual learning setup, where generative distillation is used to prevent catastrophic forgetting by distilling the entire reverse process from a teacher model to a student model learning a new task [21]. This approach has also been adopted for unlearning with SFD [7], which utilizes a distillation loss to guide the unlearning process away from an undesirable concept.

**Continual Unlearning** Real-world applications often require handling sequential removal requests, a challenge known as continual unlearning (CUL). The primary obstacle in CUL is catastrophic forgetting, driven by cumulative parameter drift, where sequential updates degrade model performance. As empirically demonstrated in our experiments (Section 5), naively applying one-shot methods causes each step to push the model’s weights further from their original state, leading to a rapid decline in image quality and retention [40].

### 3. Problem Formulation and Challenges

#### 3.1. Continual Unlearning: Problem Setting

We formally define the continual unlearning problem for text-to-image diffusion models. Given a pre-trained model  $\epsilon_{\theta_0}$  and a sequence of forget concept sets  $\{C_f^{(1)}, C_f^{(2)}, \dots, C_f^{(K)}\}$  arriving over time, our goal is to sequentially remove each concept while retaining all other concepts in the retain set  $C_r$ , where  $C_r \cap (C_f^{(1)} \cup C_f^{(2)} \cup \dots \cup C_f^{(K)}) = \emptyset$ . At each step  $i \in \{1, \dots, K\}$ , we seek to update the model parameters such that it no longer generates images for prompts  $c_f \in C_f^{(i)}$ , preserves the generation quality of retain prompts  $c_r \in C_r$ , and maintains overall image quality. Formally, we aim to find  $\theta_i = \text{Unlearn}(\theta_{i-1}, C_f^{(i)})$  satisfying: generated samples  $x \sim P_{\theta_i}(\cdot | c_f)$  should not contain recognizable instances of the forget concept for  $c_f \in C_f^{(i)}$  (unlearning),  $P_{\theta_i}(x | c_r) \approx P_{\theta_{i-1}}(x | c_r)$  for  $c_r \in C_r$  (retention), and generation quality evaluated by metrics such as FID should remain comparable to the original model (quality preservation). The key challenge is achieving this over  $K$  sequential steps without catastrophic collapse.

#### 3.2. Why Existing One-shot Methods Fail in CUL

We now analyze why existing unlearning methods, designed for single-step deletion, fail when applied sequentially in a continual unlearning setting.

##### 3.2.1. Foundations of Diffusion Models

Text-to-image diffusion models such as Stable Diffusion operate in the latent space of a pre-trained VAE [26]. Given an image  $x_0$ , the VAE encoder maps it to latent representation  $z_0 = \text{VAE}_{\text{enc}}(x_0)$ . The forward diffusion process adds Gaussian noise to  $z_0$  over  $T$  timesteps according to schedule  $\{\beta_t\}_{t=1}^T$ . With  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ :

$$q(z_t | z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t) \mathbf{I}). \quad (1)$$

The reverse process, parameterized by  $\epsilon_\theta$ , minimizes:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, z_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, c)\|_2^2], \quad (2)$$

where  $c$  is the text condition. DDIM [29] accelerates sampling via deterministic updates:

$$\begin{aligned} z_{t-1} &= \sqrt{\bar{\alpha}_{t-1}} \left( \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t, c)}{\sqrt{\bar{\alpha}_t}} \right) \\ &\quad + \sqrt{1 - \bar{\alpha}_{t-1}} \cdot \epsilon_\theta(z_t, t, c). \end{aligned} \quad (3)$$

After denoising to  $z_0$ , the final image is obtained via the VAE decoder:  $x_0 = \text{VAE}_{\text{dec}}(z_0)$ .

### 3.2.2. Typical One-shot Unlearning Approach

Existing methods [8, 9, 18, 34] typically combine two components. For effective unlearning, they employ an unlearning loss that maps the forget concept to a fixed anchor, such as an empty string [9, 34], a general [10, 18], or a random concept [8]. To preserve performance on non-target concepts, they may include a preservation term, such as a training loss on retain concepts or a KL divergence constraint to keep the model from drifting too far from its pre-unlearning distribution.

### 3.2.3. Three Key Failure Modes in Sequential Application

While effective erasure is the primary objective of unlearning, the continual setting introduces critical challenges under sequential application. When applied sequentially, existing one-shot methods trigger a cascade of degradations rather than simple additive errors. We identify three distinct yet interacting failure modes that fundamentally hinder CUL.

**Failure Mode 1: Retention Collapse (Catastrophic Forgetting).** Existing methods employ only weak preservation mechanisms (e.g., KL-based similarity constraint or placeholder mapping), which prove insufficient in the continual setting [9, 40]. With each unlearning step, the model’s knowledge of retain concepts  $C_r$  progressively degrades due to lack of explicit retention objectives, resembling the catastrophic forgetting observed in continual learning [17]. This results in global deterioration of generative capability. Over time, the model produces low-quality or nonsensical outputs even for prompts unrelated to forget concepts.

**Failure Mode 2: Compounding Ripple Effects (Collateral Damage).** A more subtle yet critical failure emerges when unlearning operations unintentionally affect semantically related or nearby concepts [9]. Even in one-shot settings, prior work has observed that mapping a target concept to a generic placeholder can unintentionally degrade related concepts, a phenomenon often referred to as a “ripple effect” [1, 40]. In a continual unlearning setup, where such mappings are applied repeatedly, we hypothesize that these ripple effects accumulate: multiple concepts may be redirected toward similar generic regions in embedding space, making it harder for the model to maintain sharp semantic boundaries between them.

**Failure Mode 3: Cumulative Parameter Drift.** Sequential unlearning steps, even when individually small, accumulate and progressively drive parameters away from their stable manifold, in line with observations from continual learning on catastrophic forgetting [17]. Each update  $\theta_i \leftarrow \theta_{i-1}$  shifts parameters farther from the original equilibrium  $\theta_0$ . Without regularization, the cumulative distance  $\|\theta_i - \theta_0\|$  increases monotonically, gradually destabilizing the model and degrading its denoising capability.

These three failure modes interact and compound across sequential steps, collectively giving rise to **concept revival** [12], where previously unlearned concepts re-emerge as the model destabilizes. As demonstrated empirically in Figure 3, existing methods collapse after 3-5 sequential unlearning steps, exhibiting severe quality degradation and concept revival. Our framework addresses each failure mode with a dedicated component, enabling stable and scalable CUL across multiple sequential deletions.

## 4. Method: Distillation-Based CUL

### 4.1. Overall Framework

We now present our distillation-based framework, which directly addresses the three failure modes identified in Section 3.2.3. We reframe each unlearning step as a multi-objective, teacher-student distillation process. The student model  $\epsilon_{\theta_i}$  is the model being trained at step  $i$ , while the previous-step teacher  $\epsilon_{\hat{\theta}_{i-1}}$  is a frozen model from step  $i-1$  that serves as a reference for what should be retained and generates synthetic data for preservation.

Our framework consists of three complementary components, each addressing a specific failure mode. *Contextual trajectory re-steering* ( $\mathcal{L}_{\text{unlearn}}$ ) achieves effective unlearning while mitigating ripple effects by redirecting the generation path through context-preserving mapping. *Generative replay* ( $\mathcal{L}_{\text{retain}}$ ) counters retention collapse through explicit knowledge transfer from the teacher. *Parameter regularization* ( $\mathcal{L}_{\text{reg}}$ ) prevents cumulative parameter drift by constraining weight changes at each step.

The final objective combines all three components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{unlearn}} \mathcal{L}_{\text{unlearn}} + \lambda_{\text{retain}} \mathcal{L}_{\text{retain}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \quad (4)$$

### 4.2. Prompt Design and Data Flow

Before detailing each component, we clarify the types of prompts used and their roles in our framework. **Forget prompts** ( $c_f \in \mathcal{D}_{\text{forget}}$ , where  $\mathcal{D}_{\text{forget}}$  is a dataset containing prompts from the forget concept set  $C_f^{(i)}$ ) contain the concept to unlearn (e.g., “Pikachu rides a horse on trail”). **Mapping prompts** ( $c_m \in \mathcal{D}_{\text{map}}$ ) are target prompts for contextual trajectory re-steering, generated using the mapping strategies detailed in Section 4.3. **Retain prompts** ( $c_r \in \mathcal{D}_{\text{retain}}$ , where  $\mathcal{D}_{\text{retain}}$  is a dataset containing prompts from the retain concept set  $C_r$ ) are for concepts to preserve, including both related and random concepts (e.g., “Squirtle figurine on office desk” is related to Pikachu).

At step  $i$ , the data flow proceeds as follows: the teacher generates images  $z_0 \sim \text{DDIM}(\epsilon_{\hat{\theta}_{i-1}}, c)$ , then noise is added  $z_t \leftarrow \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$  (with superscripts  $z_t^u$  and  $\epsilon^u$  for unlearning,  $z_s^r$  and  $\epsilon^r$  for retention to distinguish different sampling processes), and finally the student learns via distillation losses.

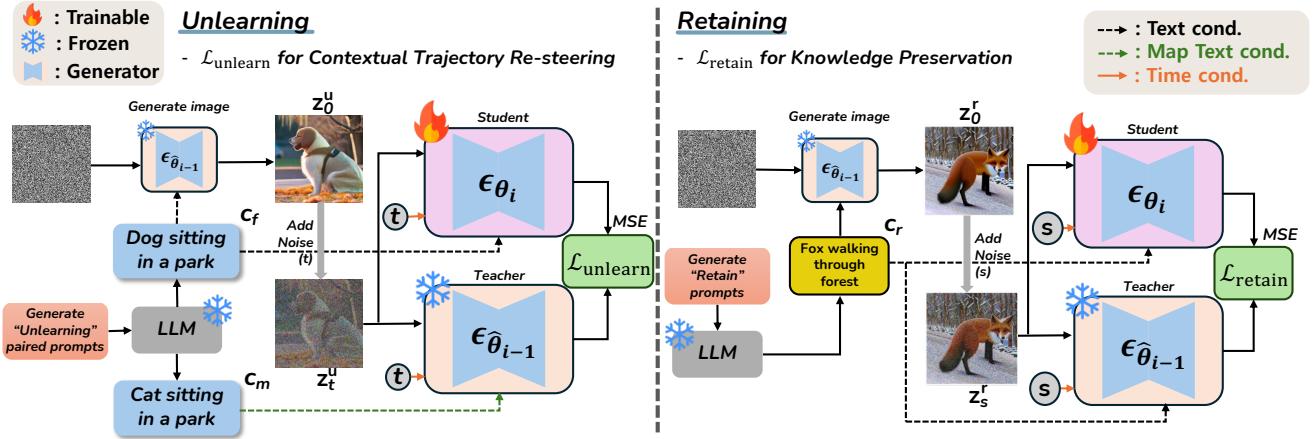


Figure 2. Visual overview of our full pipeline for both unlearning (left) and retaining (right). In the unlearning stage, an LLM generates two types of text conditions: forget prompts ( $c_f$ ) and their mapping prompts ( $c_m$ ). These are encoded using the text encoder to obtain text embeddings, and fed into a frozen teacher diffusion model ( $\epsilon_{\hat{\theta}_{i-1}}$ ) to synthesize a clean latent  $z_0^u$ , which is perturbed by the noise scheduler to produce  $z_t^u$ . The frozen teacher model is conditioned on  $z_t^u$ , the timestep  $t$ , and the mapping prompt embedding, while the trainable student model ( $\epsilon_{\theta_i}$ ) is conditioned on  $z_t^u$ ,  $t$ , and the forget prompt embedding. The unlearning objective ( $\mathcal{L}_{\text{unlearn}}$ ) is achieved by minimizing the MSE loss between the noise predictions of the teacher and the student. The retaining stage follows a similar distillation process, where the frozen teacher generates images conditioned on retain prompts ( $c_r$ ) to produce latents  $z_s^r$ , and both the teacher and student models are conditioned on the same retain prompt at timestep  $s$ , with the retention objective ( $\mathcal{L}_{\text{retain}}$ ) minimizing their prediction difference.

### 4.3. Component 1: Contextual Trajectory Re-steering for Unlearning

This component addresses **effective unlearning** and **compounding ripple effects**. Our method reframes unlearning as a targeted **contextual trajectory re-steering**. We teach the model to respond to a “forget” prompt as if it had received a “mapping” prompt, avoiding destructive parameter updates that cause ripple effects. As detailed in Algorithm 1, the student  $\epsilon_{\theta_i}$  is trained to match the teacher’s noise prediction on forget-conditioned latents, with the teacher conditioned on mapping prompts. This surgically modifies specific trajectory paths without broadly disrupting the parameter space, mitigating collateral damage to related concepts.

#### Algorithm 1 $\mathcal{L}_{\text{unlearn}}$ for Contextual Trajectory Re-steering

**Require:** Student  $\epsilon_{\theta_i}$ , Teacher  $\epsilon_{\hat{\theta}_{i-1}}$ , Paired prompts  $(\mathcal{D}_{\text{forget}}, \mathcal{D}_{\text{map}})$

- 1: Sample  $(c_f, c_m) \sim (\mathcal{D}_{\text{forget}}, \mathcal{D}_{\text{map}})$
- 2:  $z_0^u \sim \text{DDIM}(\epsilon_{\hat{\theta}_{i-1}}, c_f)$
- 3: Sample  $t \sim \mathcal{U}(1, T)$ ,  $\epsilon^u \sim \mathcal{N}(0, \mathbf{I})$
- 4:  $z_t^u \leftarrow \sqrt{\bar{\alpha}_t} z_0^u + \sqrt{1 - \bar{\alpha}_t} \epsilon^u$
- 5:  $\epsilon_{\text{student}}^u \leftarrow \epsilon_{\theta_i}(z_t^u, t, c_f)$
- 6:  $\epsilon_{\text{teacher}}^u \leftarrow \epsilon_{\hat{\theta}_{i-1}}(z_t^u, t, c_m)$
- 7:  $\mathcal{L}_{\text{unlearn}} \leftarrow \|\epsilon_{\text{teacher}}^u - \epsilon_{\text{student}}^u\|_2^2$

**Mapping Strategies.** Our contextual trajectory re-steering

method explores two mapping strategies. Many existing methods map a forget concept to a single, fixed placeholder (like “an object” or an empty string), which ignores the other benign parts of the prompt and can result in incoherent or low-quality generations. Our strategies, by contrast, use an LLM to preserve the prompt’s original context. *Fixed-Context Mapping*: In this strategy, we pre-define a fixed surrogate concept (e.g., Dog → Cat). However, instead of a simple word replacement, we use an LLM to rewrite the entire prompt to be contextually coherent with the new concept. For example, a prompt like “A dog playing with its puppies” would be intelligently mapped to “A cat playing with its kittens,” preserving the surrounding contextual information. *Adaptive-Context Mapping*: This strategy moves closer to the goal of perfect unlearning, which is to make the model behave as if it had never seen the forget concept. We provide the full forget prompt to an LLM and ask it to replace the forget concept with the *most suitable* contextual alternative. For example, as shown in our results (Figure 1), “Lionel Messi greets cheering supporters” is mapped to “A child greets cheering supporters,” while “A banana beside a fish tank” is mapped to an empty scene with just the fish tank. This demonstrates the model’s ability to find the best possible replacement for a given context.

### 4.4. Component 2: Generative Replay with Knowledge Distillation

This component addresses **retention collapse**. To counter retention collapse, we employ generative replay combined

with knowledge distillation. As detailed in Algorithm 2, the teacher generates synthetic images for retain prompts, which are noised to random timesteps. The student is trained to mimic the teacher’s denoising behavior, distilling the entire generative trajectory rather than just final outputs. This enables active knowledge consolidation at each step, preventing error accumulation. Unlike ESD [9] or UCE [10], our method shows substantially more robust retention across multiple unlearning steps.

---

**Algorithm 2**  $\mathcal{L}_{\text{retain}}$  for Knowledge Preservation

---

**Require:** Student  $\epsilon_{\theta_i}$ , Teacher  $\epsilon_{\hat{\theta}_{i-1}}$ , Retain prompts  $\mathcal{D}_{\text{retain}}$

- 1: Sample  $c_r \sim \mathcal{D}_{\text{retain}}$
- 2:  $z_0^r \sim \text{DDIM}(\epsilon_{\hat{\theta}_{i-1}}, c_r)$
- 3: Sample  $s \sim \mathcal{U}(1, T)$ ,  $\epsilon^r \sim \mathcal{N}(0, \mathbf{I})$
- 4:  $z_s^r \leftarrow \sqrt{\alpha_s} z_0^r + \sqrt{1 - \alpha_s} \epsilon^r$
- 5:  $\epsilon_{\text{student}}^r \leftarrow \epsilon_{\theta_i}(z_s^r, s, c_r)$
- 6:  $\epsilon_{\text{teacher}}^r \leftarrow \epsilon_{\hat{\theta}_{i-1}}(z_s^r, s, c_r)$
- 7:  $\mathcal{L}_{\text{retain}} \leftarrow \|\epsilon_{\text{teacher}}^r - \epsilon_{\text{student}}^r\|_2^2$

---

#### 4.5. Component 3: Parameter Regularization

This component addresses **cumulative parameter drift**. To prevent cumulative parameter drift, we apply an  $\ell_2$  regularization penalty:

$$\mathcal{L}_{\text{reg}} = \|\theta_i - \hat{\theta}_{i-1}\|_2^2. \quad (5)$$

By the triangle inequality,  $\|\theta_i - \theta_0\| \leq \sum_{j=1}^i \|\theta_j - \theta_{j-1}\|$ , so regularizing each step helps indirectly control overall deviation from the original model. This can be viewed as a simplified approximation of Elastic Weight Consolidation [17]. This term is critical for the continual setup, as it ensures only minimal parameters are affected during each unlearning step. By preventing significant parameter drift, this loss helps avoid the cumulative model degradation that often causes catastrophic failure in sequential unlearning tasks.

#### 4.6. Training Procedure

For each unlearning step  $i = 1, \dots, K$ , we initialize  $\theta_i \leftarrow \theta_{i-1}$  and freeze the teacher parameters  $\hat{\theta}_{i-1}$  (corresponding to  $\epsilon_{\hat{\theta}_{i-1}}$ ). At each training iteration, we compute all three losses and optimize the total objective (Eq. 4) using Adam optimizer. Hyperparameters are provided in Section 5.

### 5. Experiments

To validate the efficacy of our proposed continual unlearning framework, we conduct a series of experiments designed to rigorously assess both the removal of targeted knowledge and the preservation of overall model utility. This section details our experimental setup, the datasets and

prompts used, and the comprehensive suite of metrics employed for evaluation. The code is available at <https://github.com/SonyResearch/DFR-ConUnl>.

#### 5.1. Experimental Settings

**Base Model and Training.** All experiments are performed on the Stable Diffusion v1.5 model, which is the commonly used base model for experimentation (in the supplementary Section 8.6). During each unlearning and preservation step, the training process samples timesteps  $t$  and  $s$  uniformly from the range 0 to 600 (More details in supplementary section 10). The model is trained using the Adam optimizer with a learning rate of  $5 \times 10^{-6}$ . The weights for our final loss objective are set as follows:  $\lambda_{\text{unlearn}} = 1.0$ ,  $\lambda_{\text{retain}} = 10.0$ , and  $\lambda_{\text{reg}} = 0.0001$ .

**Prompt Generation and Mapping.** To construct the datasets for unlearning and preservation, we utilized the OpenAI ChatGPT API. This allowed for the generation of diverse and semantically rich prompts.

- **Forget and Neutral Prompts:** For each concept targeted for unlearning, we generated 100 unique prompts ( $\mathcal{D}_{\text{forget}}$ ). A corresponding set of 100 neutral prompts ( $\mathcal{D}_{\text{neutral}}$ ) was also generated to serve as redirection targets. We use both Fixed-Context Mapping and Adaptive-Context Mapping.
- **Retain Prompts:** A broad set of 150 prompts ( $\mathcal{D}_{\text{retain}}$ ) covering a wide range of semantically related and random concepts was generated to be used in the preservation step, ensuring the model does not suffer from catastrophic forgetting.

**Unlearning Targets.** We evaluated our method by sequentially unlearning a diverse set of 10 concepts, comprising objects, artistic styles, and famous individuals. The concepts in order are: *Pikachu*, *Brad Pitt*, *Dog*, *Golf Ball*, *Van Gogh Style*, *Apple*, *Spiderman*, *Lionel Messi*, *Cartoon Style* and *Banana*.

#### 5.2. Evaluation Metrics

We employ a multi-faceted evaluation strategy to provide a holistic view of the model’s performance, focusing on unlearning efficacy, knowledge retention, and overall generative quality. Our evaluation framework is not arbitrary; it is built upon and extends the methodologies established in recent unlearning benchmarks. Our core metrics for unlearning efficacy and general knowledge preservation are inspired by benchmarks like UnlearnCanvas [40], from which we adapt Unlearning Accuracy (UA) and General Retention Accuracy (GRA). To more precisely quantify the “ripple effect”, a critical failure mode in sequential unlearning, we also integrate metrics inspired by EraseBench [1]. This includes evaluating both the accuracy Related Retain Ac-

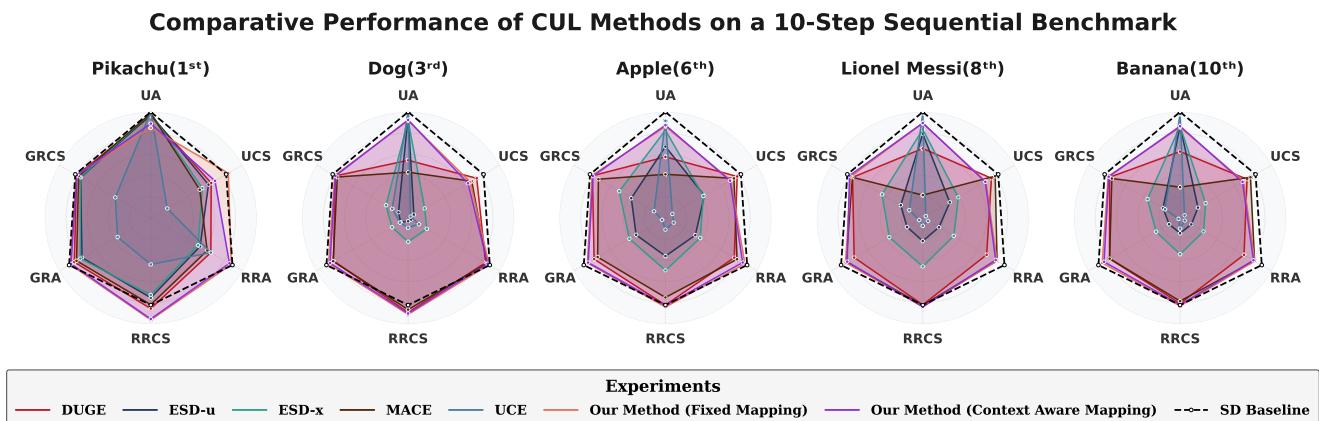


Figure 3. A visual summary of our method’s performance (both Fixed Context and Adaptive Context mapping) against SOTA baselines across 10 sequential unlearning steps. Here the ‘SD Baseline’ indicates performance of base model SD v1.5 for easier comparison, here we assume UA be 100 and for rest of the metrics to be overall average CS and Accuracy. The plot starkly illustrates the “retention collapse” of methods like ESD-x, which shrink to the center, and the unlearning failure of methods like UCE and MACE. Our methods are the only ones that maintain a large, stable shape, indicatin[g a successful balance of all criteria. Metrics shown are: Unlearning Accuracy (UA), Unlearning CLIP Score (UCS), Related Retention Accuracy (RRA), Related Retention CLIP Score (RRCS), General Retention Accuracy (GRA), and General Retention CLIP Score (GRCS). A detailed breakdown of these metrics is provided in the supplementary material Table 5.

curacy and CLIP score for semantically related concepts, allowing for a more granular assessment of collateral damage.

**Unlearning and Retention Accuracy.** To automate and standardize evaluation, we use the Qwen2.5-VL-7B-Instruct model [2] as our evaluator. For each metric, we generate images and ask a binary question (e.g., “Does this image contain a dog?”) to compute accuracy. In contrast to ImageNet-style classifiers or UnlearnCanvas heads, which are restricted to narrow label spaces and struggle with celebrities, artistic styles, and other diverse concepts, modern VLMs like Qwen2.5-VL proved substantially more reliable. By comparing conflicting predictions between these baselines and the VLM, we consistently found the VLM to yield the correct interpretation, and we include supporting evidence for this observation in the supplementary material. For evaluation, we use diverse prompts that are not seen in the  $\mathcal{D}_{\text{forget}}$  and  $\mathcal{D}_{\text{retain}}$  sets used during continual unlearning, in order to ensure robustness to phrasing variations and to avoid prompt memorization.

- **Unlearning Accuracy (UA) & CLIP Score (UCS):** UA measures the effectiveness of the forgetting process. We generate images using evaluation prompts and query the VLM to determine if the unlearned concept is present; a high UA (a high percentage of “No” answers) indicates successful unlearning. We also measure the UCS for these same prompts. A high UCS is critical as it demonstrates “in-prompt retainability” [25], ensuring the model correctly generates the benign parts of the prompt even

after the target concept is removed. So the ideal outcome is both a high UA and a high UCS.

- **Related Retention (RRA) & CLIP Score (RRCS):** RRA measures the unintended impact on semantically adjacent concepts (i.e., the “ripple effect” [1, 40]). We evaluate the model’s ability to generate concepts closely related to the unlearned one (e.g., testing “Impressionism Style” after unlearning “Van Gogh Style”). A high RRA indicates surgical unlearning, and RRCS ensures the text-to-image alignment for these concepts was not degraded.
  - **General Retention Accuracy (GRA) & CLIP Score (GRCS):** GRA assesses overall knowledge preservation by testing the model on a broad set of unrelated prompts to ensure it has not suffered from catastrophic forgetting. Similarly, the GRCS verifies that the model’s text-to-image alignment for this broad set of general concepts remains intact.

**Generative Quality:** Beyond accuracy and text-alignment, we follow standard practice to evaluate the fundamental quality of the generated images. We adopt the Fréchet Inception Distance (FID) to evaluate image quality and diversity, following the methodology used in the UnlearnCanvas benchmark.

### **5.3. Experimental Results and Analysis**

We have conducted extensive evaluations to assess the effectiveness of our multi-objective distillation framework against the Continual Unlearning setup.

**How well do the existing Unlearning methods perform in Continual Unlearning setup?** Based on Figure 3, we

can see that existing unlearning methods are fundamentally ill-equipped for the sequential demands of a continual unlearning (CUL) setup. When applied repeatedly, they suffer “catastrophic forgetting” and fail in distinct, opposing ways. Methods involving broad fine-tuning, like ESD-u and ESD-x, suffer from catastrophic “retention collapse.” The core flaw is that each unlearning step compounds parameter drift, triggering a “cascade of degradation” that destroys the model’s general knowledge and results in a total breakdown of generative quality.

Conversely, methods designed to be more surgical to avoid this collapse fail at the primary unlearning task. UCE, a training-free method, is “too minimal”; it perfectly preserves the model but is fundamentally ineffective at unlearning from the start. MACE successfully isolates changes in parameter-efficient LoRA modules, but its unlearning efficacy progressively degrades as these minimal, isolated edits accumulate and conflict over subsequent tasks. Finally, DUGE, while designed for CUL, offers an “ineffective compromise.” Its mechanism to prevent “generalization erosion” is “over-cautious,” leading to poor unlearning performance while still allowing significant “ripple effects” and knowledge drift to accumulate over time.

#### Does our work improve Unlearning in Continual setup?

Yes. Fig. 3 shows that our methods shows consistent improvements in all the aspects towards satisfying the desiderata for Unlearning.

**How does mapping or surrogate concept affect the Unlearning process?** Both our mapping strategies, Fixed-Context Mapping and Adaptive-Context Mapping, perform well and, as shown in our 10-step sequential benchmark, are far more robust than SOTA baselines which suffer from “retention collapse” or unlearning failure. The performance between our two methods is very close, but they reveal a critical trade-off. The Adaptive-Context Mapping (e.g., “Dog playing” → “Kids playing”) achieves more effective unlearning, as evidenced by its superior Unlearning Accuracy (more results in supplementary). This aligns with recent findings suggesting that erasure is a local operation in the concept space; by dynamically finding the optimal, semantically closest target for each specific context, this strategy performs a more minimal and surgically precise edit in the latent space [6]. However, this precision comes at a small cost. The Fixed-Context Mapping (e.g., “Dog” → “Cat”) provides superior knowledge retention, with consistently higher RRA and GRA scores. We hypothesize this is because the “trajectory re-steering” objective ( $\mathcal{L}_{\text{unlearn}}$ ) is highly stable; every “forget” prompt is re-steered toward the same fixed region in the latent space. This consistency creates a strong, stable anchor that minimizes the “ripple effect” on related concepts, whereas the added variance of a moving target in the adaptive context, while more precise for unlearning, has a slightly larger impact on retention.

**How does our method achieve this, and what is the contribution of each objective?** To quantify the role of each component, we run a comprehensive ablation study, with results summarized in Table 1. All metrics are reported after 10 sequential steps of CUL. We analyze the effects of using our contextual trajectory re-steering loss ( $\mathcal{L}_{\text{unlearn}}$ ) alone, and in combination with the generative replay loss ( $\mathcal{L}_{\text{retain}}$ ) and the parameter regularization loss ( $\mathcal{L}_{\text{reg}}$ ).

Method	UA ↑	UCS ↑	RRA ↑	RRCS ↑	GRA ↑	GRCS ↑
Original SD	–	–	–	–	0.89	32.8
$\mathcal{L}_{\text{unlearn}}$ only	0.94	27.1	0.28	27.7	0.56	29.1
$\mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{retain}}$	0.95	28.0	0.65	31.2	0.75	31.1
$\mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{reg}}$	0.82	30.3	0.59	31.2	0.74	31.3
<b>Ours (Full Model)</b>	<b>0.86</b>	<b>30.4</b>	<b>0.81</b>	<b>33.0</b>	<b>0.85</b>	<b>32.1</b>

Table 1. Ablation study on the components of our framework after 10 continual unlearning steps. We analyze the contribution of the generative replay loss ( $\mathcal{L}_{\text{retain}}$ ) and the parameter regularization loss ( $\mathcal{L}_{\text{reg}}$ ). The full model demonstrates the necessity of both stability components to balance unlearning (high UA) and prevent retention collapse (high RRA/GRA).

Our analysis of the results reveals the distinct role each component plays:

- **$\mathcal{L}_{\text{unlearn}}$  only:** Using only the trajectory re-steering loss is insufficient for the CUL setting. While effective at unlearning, it suffers from a severe “compounding ripple effect” and “retention collapse,” evident by the catastrophic drop in all retention scores.
- **$\mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{retain}}$ :** Adding the generative replay loss ( $\mathcal{L}_{\text{retain}}$ ) directly counters retention collapse, providing the primary mechanism for functional stability. This component forces the student to mimic the teacher’s behavior on retained concepts, dramatically improving retention.
- **$\mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{reg}}$ :** Adding only the parameter regularization loss ( $\mathcal{L}_{\text{reg}}$ ) provides parameter stability and helps prevent cumulative parameter drift. While this improves retention, it also makes the model more resistant to change, which slightly hinders the unlearning process itself.
- **Ours (Full Model):** The full framework demonstrates the crucial synergy between the two stability components. The generative replay ( $\mathcal{L}_{\text{retain}}$ ) provides functional stability, while the parameter regularization ( $\mathcal{L}_{\text{reg}}$ ) provides the long-term parameter stability needed to prevent cumulative drift over 10 steps. This combination achieves the best overall balance, successfully mitigating both “retention collapse” and “parameter drift” while achieving the highest overall retention scores.

## 6. Conclusion

As real-world data removal requests for privacy or copyright reasons arrive sequentially, this paper introduces a distillation-based framework for continual unlearning (CUL) in text-to-image diffusion models, designed to

overcome the catastrophic forgetting and quality collapse that plagues existing methods. Our solution successfully addresses this challenge by decomposing the objective into three components: (1) contextual trajectory re-steering ( $\mathcal{L}_{\text{unlearn}}$ ) to surgically redirect and unlearn target concepts, (2) generative replay with distillation ( $\mathcal{L}_{\text{retain}}$ ) to preserve the model’s behavior for all retained concepts, and (3) parameter regularization ( $\mathcal{L}_{\text{reg}}$ ) to prevent cumulative weight drift. As demonstrated qualitatively in our 10-step sequential unlearning experiments, our method effectively removes targeted concepts while maintaining high generative fidelity for retained concepts, in stark contrast to baseline methods, which suffer from severe quality degradation and collapse into noise.

## References

- [1] Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najuong Kim, Junfeng He, Cristina Nader Vasconcelos, Deepak Ramachandran, Golnoosh Farnadi, Katherine Heller, Mohammad Havaei, and Negar Rostamzadeh. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2, 3, 4, 6, 7
- [2] 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. 7, 6
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. 2
- [4] Shristi Das Biswas, Arani Roy, and Kaushik Roy. CURE: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3
- [5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 2
- [6] Anh Tuan Bui, Thuy-Trang Vu, Long Tung Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. In *The Thirteenth International Conference on Learning Representations*, 2025. 8
- [7] Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [8] 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 *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4
- [9] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023. 1, 2, 3, 4, 6
- [10] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2, 3, 4, 6
- [11] Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025. 2
- [12] Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13393–13402, 2025. 2, 4
- [13] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36: 17170–17194, 2023. 2
- [14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [15] Chia Yi Hsu, Yu Lin Tsai, Chulin Xie, Chih Hsun Lin, Jia You Chen, Bo Li, Pin Yu Chen, Chia Mu Yu, and Chun Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *12th International Conference on Learning Representations*, 2024. 3
- [16] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024. 3
- [17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 4, 6
- [18] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2, 3, 4
- [19] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 1, 3
- [20] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 3
- [21] Sergi Masip, Pau Rodríguez, Tinne Tuytelaars, and Gido M. van de Ven. Continual learning of diffusion models with generative distillation. In *COLLAs*, pages 431–456, 2024. 3
- [22] Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4211–4221, 2025. 3
- [23] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. Red-teaming the stable diffusion safety filter. In *NeurIPS ML Safety Workshop*, 2022. 2
- [24] Protection Regulation. General data protection regulation. *Intouch*, 25:1–5, 2018. 2
- [25] Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for text-to-image diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28769–28778, 2025. 7, 6
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [27] Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2966–2975, 2025. 3
- [28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 3
- [29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3
- [30] Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23765–23774, 2025. 3
- [31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricu, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2
- [32] Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Continual unlearning for foundational text-to-image models without generalization erosion. *arXiv preprint arXiv:2503.13769*, 2025. 1, 2
- [33] Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pages 367–384. Springer, 2024. 3
- [34] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28263–28273, 2025. 2, 4
- [35] Abdukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual LLM unlearning. In *Forty-second International Conference on Machine Learning*, 2025. 2
- [36] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 2
- [37] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1755–1764, 2024. 3
- [38] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024. 3
- [39] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2024. 3
- [40] Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2, 3, 4, 6, 7

# Distill, Forget, Repeat: A Framework for Continual Unlearning in Text-to-Image Diffusion Models

## Supplementary Material

### 7. Overview

In this supplementary material, we provide additional details and results to support the findings in the main paper. The content is organized as follows:

- Section 8: Implementation details.
- Section 9: Evaluation.
- Section 10: Ablation Study with Different T.
- Section 11: Main quantitative results.
- Section 12: Additional qualitative visualizations.

### 8. Implementation Details

#### 8.1. Synthetic Data Generation for Continual Unlearning

To enable efficient continual unlearning without the computational overhead of full image generation, we implement a latent-space generative replay mechanism. Unlike standard approaches that generate pixel-space images ( $x_0$ ) and then re-encode them using a VAE, our Previous-Step Teacher  $\epsilon_{\hat{\theta}_{t-1}}$  generates *clean latents* ( $z_0$ ) directly.

We use the Denoising Diffusion Implicit Models (DDIM) sampler to synthesize these latents. Specifically, for a given prompt  $c$ , the teacher model produces a trajectory in the latent space, yielding a final clean latent  $z_0$ . This latent is then directly perturbed with Gaussian noise to creating the training targets  $z_t$  for the Student model. By bypassing the VAE’s decoder and encoder steps entirely, we significantly reduce memory usage and training time while maintaining the statistical fidelity of the replay data.

#### 8.2. Prompt Generation

To ensure robustness and semantic diversity, we construct our datasets using the OpenAI GPT-5 API. We employ distinct system prompts to guide the generation of “Forget”, “Neutral”, and “Retain” prompts, ensuring they adhere to specific visual grounding and diversity constraints.

##### 8.2.1. Prompt Generation for Unlearning

For the unlearning objective, we generate paired data ( $c_{\text{forget}}, c_{\text{neutral}}$ ). We utilize two distinct strategies, each governed by a specific system prompt fed to the LLM.

**Strategy 1: Adaptive-Context Mapping.** In this strategy, the LLM is instructed to find the most semantically plausible replacement for the target concept within the given context.

#### Adaptive Context Mapping - System Prompt

You are a dataset prompt writer for diffusion models (e.g., Stable Diffusion). Your job is to produce short, simple, natural-language prompts that are easy for diffusion models to parse and that depict a clearly visible instance of the specified concept in-frame.

##### GENERAL STYLE

- 4-10 words per prompt.
- One simple main clause; present tense preferred.
- Everyday vocabulary only; 0-1 adjectives
- .
- No tags/parameters (no colons, aspect ratios, seeds, CFG, hashtags).
- Minimal punctuation (commas/periods only when needed).

##### VISUAL GROUNDING (MANDATORY for both TRAIN and VALIDATION)

- The concept **MUST** be a **visible, concrete subject** in the image.
- If the concept is polysemous, use the **physical object** sense (e.g., "apple" = fruit, **not** the company).
- The scene must make the concept **clearly depictable and in-frame**. Prefer contexts where the subject is visible (e.g., not hidden, microscopic, or distant).
- Match number and determiners consistently (a/an/one/two/many).

##### MAPPED COUNTERPARTS (CONTEXT-PRESERVING REPLACEMENT OR REMOVAL)

You must produce a "mapped" counterpart for each training prompt by **replacing the specified concept** with a **context-appropriate, visually plausible subject**, while preserving the sentence’s structure and scene.

- If a **natural replacement** makes the prompt more realistic or coherent (even if it changes category, e.g., object  $\rightarrow$  person), prefer that.
- If no sensible replacement exists, **REMOVE** the concept tokens and keep

- the rest intact.
- The mapped prompt must describe a \*\* visually depictable\*\* subject or scene (no abstractions like "happiness" or "freedom").
  - Preserve number/determiners and the grammatical shape of the sentence.
  - The mapped prompt must remain \*\* realistic and visually grounded\*\* - not surreal or symbolic.

---

CONCEPT TYPES & MAPPING RULES

---

- 1) FLEXIBLE REPLACEMENT RULE (Default)
  - Choose a replacement that fits the \*\* scene context\*\*, even if it's from a different semantic category.

Example:  
"An apple sitting on a beach" -> "A person sitting on a beach"  
(cross-category but visually coherent)
- 2) SAME-CATEGORY REPLACEMENT (When Naturally Fits)
  - Replace with a category sibling when it makes contextual sense.

Example:  
"A dog chasing a ball" -> "A cat chasing a ball"
- 3) PERSON OR NAMED ENTITY
  - Replace with a neutral, visually grounded role or human activity.

Example:  
"Barack Obama giving a speech" -> "A teacher giving a speech"
- 4) ARTISTIC STYLE OR ARTIST NAME
  - If the concept refers to a style or artist, either:
    - a) DROP the style phrase but keep the rest of the scene, or
    - b) SWAP to a generic style label like "cartoon style" or "realistic photo".

Example:  
"A landscape in Van Gogh style" ->  
"A landscape in cartoon style"
- 5) SAFETY-SENSITIVE CONTENT (weapons, nudity, harm)
  - All training prompts must be \*\*non-graphic, non-violent, and legal\*\*.
  - Mapped version should replace with a \*\*benign analogue\*\* (e.g., "knife" -> "spoon") or remove phrase entirely.

---

DIVERSITY REQUIREMENTS

---

- Use \*\*at least 5 distinct replacement concepts\*\* across the entire mapped dataset.

- Replacement concepts should vary in category (person, animal, object, etc.)
- Ensure contextual appropriateness - avoid trivial visual similarity as the only reason for mapping.
- Across prompts, balance:
  - Indoor vs. outdoor
  - Day vs. night
  - Weather conditions
  - Simple everyday activities (walking, cooking, reading, sitting, playing, riding, shopping, etc.)
  - Solitary vs. small-group scenes
  - Near vs. far framing

---

VALIDATION SET (20 PROMPTS)

---

- Create 20 \*\*distinct\*\* simple prompts containing the concept.
- Each must satisfy:
  - 4-8 words
  - Natural, realistic, everyday phrasing
  - The concept must be \*\*visible in-frame\*\*
- No overlaps or trivial paraphrases of training prompts.

---

OUTPUT FORMAT (STRICT JSON)

---

Return a single JSON object with the following structure:

```
{
  "concept": "<string, as given>",
  "train_count": <int N>,
  "train_pairs": [
    {
      "id": <int, 1-based>,
      "src": "<short prompt including the concept>",
      "mapped": "<short prompt with concept replaced or removed>",
      "mapping_type": "<'replace' | 'remove'>",
      "note": "<very short rationale for the chosen mapped concept>"
    },
    ...
  ],
  "validation_count": 20,
  "validation_prompts": [<short prompt>,
    ...
  ]
}
```

**Strategy 2: Fixed Context Mapping.** In this strategy, the target concept is mapped to a pre-defined surrogate (e.g., Dog → Cat) while rewriting the sentence to maintain grammatical and semantic coherence.

## Fixed Context Mapping - System Prompt

You are a dataset prompt writer for diffusion models (e.g., Stable Diffusion). Your job is to produce short, simple, natural-language prompts that are easy for diffusion models to parse and that depict a clearly visible instance of the specified concept in-frame.

### GOAL

Produce a JSON dataset where each TRAIN pair has:

- "src": a short natural prompt that visibly depicts the INPUT\_CONCEPT.
- "mapped": the SAME scene but with INPUT\_CONCEPT replaced by MAPPED\_CONCEPT.
- "mapping\_type": "replace" in almost all cases; use "remove" ONLY if a literal replacement would be grammatically impossible, visually implausible, or unsafe.

### GENERAL STYLE

- 5-12 words, present tense, everyday vocabulary, <=1 comma, <=2 adjectives.
- No tags/parameters (no aspect ratios, seeds, CFG, hashtags).
- Minimal punctuation; keep sentences simple and independent.

### VISUAL GROUNDING

- INPUT\_CONCEPT must be a visible, concrete subject in "src".
- The "mapped" sentence must remain depictable and keep the same scene framing and action.

### FIXED MAPPING RULES (MANDATORY)

- INPUT\_CONCEPT is given as CONCEPT.
- MAPPED\_CONCEPT is given as MAPPED\_CONCEPT.
- Replace INPUT\_CONCEPT tokens with MAPPED\_CONCEPT while preserving:
  - \* Number/plurality ("a dog" -> "a cat", "two dogs" -> "two cats").
  - \* Determiners and articles ("a/an/one/two/many"); fix a/an as needed.
  - \* Core syntax and structure of the sentence.
- The INPUT\_CONCEPT might have different names and types also try to replace them with MAPPED\_CONCEPT. (For example, if INPUT\_CONCEPT is "dog", then "puppy", "canine" or "German Shepherd" in the prompts. All these variations should be replaced with the corresponding form of MAPPED\_CONCEPT, such as "a kitten", "cat", or types of cat.)
- Do NOT invent a different replacement; always use exactly MAPPED\_CONCEPT if sensible.

- If replacement would be nonsensical/unsafe (rare), set mapping\_type="remove" and drop the concept phrase; keep the rest intact.

### CONCEPT TYPES

- 1) OBJECT/ANIMAL: replace with MAPPED\_CONCEPT as the new subject. Keep count and determiners aligned.
- 2) PERSON (generic or named): keep neutral content. If MAPPED\_CONCEPT is a named person, use the name neutrally; if generic (e.g., "a person", "a runner"), use it verbatim.
- 3) ARTISTIC STYLE: if INPUT\_CONCEPT is a style phrase (e.g., "in Van Gogh style"), swap just the style to "in MAPPED\_CONCEPT style" and keep the scene identical.
- 4) SAFETY-SENSITIVE: only neutral, non-graphic content. If literal replacement is unsafe, use mapping\_type="remove".

### DIVERSITY

- Across the N training pairs: vary indoor/outdoor, day/night, weather, actions, distance (near/far), solo vs small group.
- No near-duplicates; avoid trivial rewording.

### VALIDATION (20 prompts)

- New scenes not used in training.
- Must visibly depict the INPUT\_CONCEPT (not the mapped one).
- Same brevity and style constraints.

### OUTPUT FORMAT (STRICT JSON)

```
{
  "concept": "<string, as given>",
  "train_count": <int N>,
  "train_pairs": [
    {
      "id": <int, 1-based>,
      "src": "<short prompt including the INPUT_CONCEPT>",
      "mapped": "<short prompt where INPUT_CONCEPT is replaced by MAPPED_CONCEPT or removed>",
      "mapping_type": "<'replace' | 'remove'>",
      "note": "<very short rationale (e.g., 'dog->cat; same scene')>"
    },
    ...
  ],
  "validation_count": 20,
  "validation_prompts": ["<short prompt>",
    ...
  ]
}
```

### 8.2.2. Prompt Generation for Retaining

For the preservation set ( $\mathcal{D}_{\text{retain}}$ ), we generate diverse prompts that do not contain the target concept. We use the

following system prompts to ensure a mix of related and unrelated concepts.

### Related Concepts System Prompt

You are a helpful assistant tasked with finding related concepts. Given an input concept, provide a list of exactly 10 distinct concepts that are related to it, with good diversity:

- Include a mix of:
  1. \*\*Close or sibling items\*\* (e.g., similar type or category) (Around 4-6 items),
  2. \*\*Broader category terms\*\* (e.g., the class or group it belongs to) (2-4 items),
  3. \*\*Contextual or analogical items\*\* (things that commonly co-occur, appear together, or share thematic context) (2-4 items)
- (Very Important) Are \*\*not synonyms, subtypes, accessories, or activities that directly imply the INPUT\_CONCEPT \*\*.(for example, if INPUT is "dog", do not include "leash" or "bone" or "puppy").
- Avoid concepts which contains or strongly imply the INPUT\_CONCEPT.
- Use simple, common concepts that are easy for image generation models to understand.
- Avoid rare, obscure, technical, or brand-specific terms.
- Include broad generic concepts if relevant (e.g., "fruit", "vehicle", "person", "animal").
- If the input is a specific person or celebrity, include a few other similar famous individuals or colleagues, and also include a generic person descriptor (e.g., "a man", "a woman", or a related role).
- If the input is an artistic style or medium, include other popular art styles or mediums.
- If the input is a weapon or harmful concept, list benign alternatives in a similar context (e.g., replace weapons with cameras, replace fighting with sports).
- \*\*Output format:\*\* Return exactly 10 concepts, one per line, with no explanations or numbering.

- \*\*General Constraints (apply to all prompts):\*\*
  - Each prompt must be 5-10 words long, using simple, everyday vocabulary.
  - Use present tense and natural grammar; use <=1 comma and <=1 adjective per prompt.
  - No special tags or parameters (no ":" or "--ar 16:9" etc.).
  - \*\*Do not include the INPUT concept or any obvious synonym/nickname for it in any prompt.\*\*
- \*\*Task Distribution:\*\*

You will be given a list of 10 RELATED\_CONCEPTS (numbered 1-10) for the INPUT concept.

Create prompts as follows:

  - For each of the 10 concepts: generate \*\*10 prompts\*\* (total \*\*100\*\* prompts)
  - Then generate \*\*50 additional random prompts\*\* unrelated to the INPUT concept or its direct domain.
- \*\*Content Guidelines:\*\*
  - For prompts using the related concepts: describe simple, plausible scenes or subjects involving that concept (e.g., common actions like walking, reading, playing; or typical settings).
  - Ensure variety in scenes and wording: mix indoor/outdoor settings, day/night times, different weather, and a range of activities.
  - Do \*\*not\*\* repeat the same sentence structure across prompts or only swap the subject; avoid near-duplicates or trivial paraphrasing across all prompts.
  - For the 50 unrelated prompts: choose topics completely outside the INPUT concept's category. (They should not mention or strongly imply the INPUT concept or the related concepts.)
  - Keep all prompts \*\*safe and neutral\*\* in tone (no sexualization of real people, no graphic violence, etc.).
- \*\*Output Format:\*\*
  - Output a total of \*\*150 prompts\*\*, one prompt per line (no blank lines).
  - Do not include any numbering, bullet points, or section titles. Just provide the prompts as a continuous list.

### Prompt Generation System Prompt

You are a prompt generator for Stable Diffusion models.  
Follow these rules and tasks to produce a list of prompts:

### 8.3. Qualitative Examples of Dataset Prompts

To illustrate the diversity and behavior of our mapping strategies, we provide concrete examples of the prompts generated for the concept "Dog".

### 8.3.1. Context-Aware (Dynamic) Mapping Samples

In this strategy, the LLM dynamically selects a replacement concept that best fits the specific context of the scene (e.g., replacing a dog with a child in a playing context, or a duck in a swimming context).

Source Prompt ( $c_{\text{forget}}$ )	Mapped Prompt ( $c_{\text{neutral}}$ )	Rationale
A Dog plays with a ball	A child plays with a ball	Child fits play context
A Dog looks out the window	A cat looks out the window	Common pet behavior
A Dog swims in the lake	A duck swims in the lake	Aquatic animal replacement
A Dog rides in a car	A child rides in a car	Human passenger replacement

### 8.3.2. Fixed Mapping Samples

In this strategy, the target concept “Dog” is consistently mapped to “Cat”. As shown below, the LLM intelligently handles linguistic variations (e.g., mapping “puppy” → “kitten”) and preserves the sentence structure.

ID	Generated Prompt	Mapped	Mapping Note
1	a cat runs across a sunny park	dog→cat; same scene	
2	two cats nap on the living room rug	dogs→cats; same scene	
3	the black cat sits by the front door	dog→cat; same scene	
4	a kitten sleeps curled on a blanket	puppy→kitten; same scene	
5	a cat carries a red ball	breed→cat; same scene	
6	three cats splash in a shallow creek	dogs→cats; same scene	
7	a small cat peers out the window	dog→cat; same scene	
8	an old cat rests under the oak tree	dog→cat; same scene	
9	a spotted cat shakes off rainwater	dog→cat; same scene	
10	a cat trots along a snowy trail	breed→cat; same scene	

### 8.3.3. Validation Prompts

For evaluation, we use a held-out set of 20 prompts per concept. These are distinct from the training prompts to ensure we are testing generalization rather than memorization.

- The Dog stands on a hill
- The Dog sleeps on a blanket
- The Dog waits by the gate
- The Dog runs on the trail
- The Dog sits under the desk
- The Dog watches cars pass by
- The Dog lies on cool grass
- The Dog chews a rubber toy
- The Dog drinks from a fountain
- The Dog looks through the fence
- The Dog rides in a canoe
- The Dog rests on the porch steps
- The Dog peeks from a cardboard box
- The Dog sits by the window fan
- The Dog stands near a puddle
- The Dog trots across the field
- The Dog watches kids playing soccer
- The Dog naps beside the heater
- The Dog jumps into a pile of leaves
- The Dog sits under string lights

## 8.4. Hyperparameter Settings

All experiments are performed using the Stable Diffusion v1.5 base model. The specific hyperparameters used for our best-performing continual unlearning experiments are detailed in Table 2.

Table 2. Hyperparameter settings for Continual Unlearning.

Hyperparameter	Value
Base Model	SD v1.5
Optimizer	AdamW
Learning Rate	$1.0 \times 10^{-5}$
LR Scheduler	Constant w/ Warmup
Warmup Steps	800
Batch Size (per device)	13
Grad. Accumulation	2
Precision	FP16
Max Training Steps	600
Timestep Range ( $t$ )	500 - 1000
<b>Loss Weights</b>	
$\lambda_{\text{unlearn}}$ (Unlearning)	1.0
$\lambda_{\text{preserve}}$ (Preservation)	10.0
$\lambda_{\text{reg}}$ (Regularization)	0.0001

## 8.5. Computing Resources

All experiments were conducted on a high-performance computing cluster. Each experimental run was distributed across 3 to 4 NVIDIA H100 GPUs to parallelize the teacher

Base Model	UA $\uparrow$	UCS $\uparrow$	RRA $\uparrow$	RRCS $\uparrow$	GRA $\uparrow$	GRCS $\uparrow$
SD - 1.5	0.86	30.4	0.81	33.0	0.85	32.1
SD - 1.4	0.83	30.8	0.84	33.0	0.83	32.1
SD - 2.1 - base	0.79	30.91	0.83	33.1	0.84	32.2

Table 3. Performance across different base Stable Diffusion models. This shows that our method is able to generalize across different base models in continual 10 concept unlearning setup.

latent generation and student training processes. The implementation relies on the PyTorch framework and the Hugging Face Diffusers library.

### 8.6. Ablation Study with different base models

We conducted experiments across Stable Diffusion v2.1-base and Stable Diffusion v1.4 in the same setup, and both models showed good results with the same set of hyperparameters as shown in Table 3.

## 9. Evaluation

We employ a multi-faceted evaluation strategy to provide a holistic view of the model’s performance, focusing on unlearning efficacy, knowledge retention, and overall generative quality. Our framework extends methodologies from recent benchmarks [1, 40] to the continual setting, quantifying not just the erasure of target concepts but also the stability of the model’s distributed knowledge.

### 9.1. VLM-based Automated Evaluation

To automate and standardize the assessment of unlearning and retention, we utilize the **Qwen2.5-VL-7B-Instruct** model [2] as an impartial evaluator. For each accuracy metric, we generate a set of images and pose a closed-ended binary question to the VLM (e.g., “Does this image contain a dog?”).

We adopt this VLM-based approach to overcome the significant limitations of traditional fixed-vocabulary classifiers (e.g., ImageNet-1k pre-trained models) or the specific category heads used in UnlearnCanvas. These conventional evaluators are inherently restricted to a narrow set of predefined classes and struggle to recognize specific identities (e.g., *Lionel Messi*), complex artistic styles, and fine-grained object variations. In contrast, modern VLMs like Qwen2.5-VL possess robust open-vocabulary recognition capabilities, making them substantially more reliable for evaluating the diverse range of concepts encountered in real-world unlearning scenarios. We also manually checked and found these Classifiers were poor in classifying out of distribution and AI generated images using original Stable Diffusion models and compared to this VLM’s performed quite better.

### 9.2. Dynamic Evaluation Protocol

To ensure robust evaluation, we construct three distinct datasets ( $\mathcal{D}_{\text{eval\_forget}}$ ,  $\mathcal{D}_{\text{eval\_related}}$ ,  $\mathcal{D}_{\text{eval\_general}}$ ). Crucially, these prompts are disjoint from the training sets to test generalization and avoid memorization.

**Cumulative Evaluation Scale.** In a continual setup, evaluation must account for the growing history of unlearned concepts. We track a total of approximately **60 distinct concepts** (spanning targets, related concepts, and general themes). For every concept, we use **20 unique evaluation prompts** and generate **8 images per prompt**. This results in a comprehensive evaluation of approximately **10,000 images** for every model checkpoint.

**Evolving Related Sets.** Unlike one-shot settings, our evaluation sets evolve to rigorously test for stability and revival.

- **Cumulative Related Set:** The set of related concepts expands with each step. For instance, after unlearning *Pikachu* (Step 1), we evaluate on concepts related to Pikachu. After unlearning *Brad Pitt* (Step 2), the related set becomes the union of concepts related to Pikachu AND Brad Pitt. This cumulative tracking is vital to ensure that previously retained concepts do not degrade during later unlearning steps.
- **General Set:** This set remains fixed and disjoint, consisting of broad concepts unrelated to any of the unlearning targets. This serves as a control group to measure the “General Retention” of the model’s base capabilities.

**Related Concept Identification.** To identify semantically entangled concepts, we follow the methodology of EraseBench [1]. We use GPT-5 to generate relevant neighbors using the prompt: “Your main task is to help identify concepts for evaluating text-to-image models. The key idea is to identify 3-4 concepts that are semantically entangled with the **Given Concept...**”. This yields adjacent concepts (e.g., *Pikachu*  $\rightarrow$  *Squirtle*, *Charmander*) used to measure the ripple effect.

### 9.3. Metrics

We report the following metrics to assess the trade-off between erasure and preservation:

- **Unlearning Accuracy (UA) & CLIP Score (UCS):** UA measures the effectiveness of erasure; a high UA (high percentage of “No” answers from the VLM) indicates successful unlearning. Simultaneously, we measure the **Unlearning CLIP Score (UCS)** on these same prompts. High UCS is critical as it serves as a proxy for “in-prompt retainability” [25], ensuring that while the target concept is removed, the model still respects the benign context of the prompt. The ideal outcome is a **high UA** (concept removed) paired with a **high UCS** (context preserved).
- **Related Retention Accuracy (RRA) & CLIP Score**

Timestep Config.	UA $\uparrow$	UCS $\uparrow$	RRA $\uparrow$	RRCS $\uparrow$	GRA $\uparrow$	GRCS $\uparrow$
$T = 300$	0.82	31.8	<b>0.88</b>	<b>34.6</b>	<b>0.86</b>	<b>32.5</b>
$T = 600$ ( <b>Ours</b> )	0.91	31.0	0.86	34.2	0.85	32.4
$T = 800$	<b>0.93</b>	30.3	0.85	33.9	0.84	32.3
$T = 1000$	0.82	31.0	0.86	34.1	0.85	32.4

Table 4. **Impact of Timestep Range ( $T$ ) on Sequential Unlearning.** Results averaged over a 4-concept sequence.  $T = 600$  offers the best balance, achieving high unlearning efficacy without the instability observed at  $T = 1000$  or the poor erasure at  $T = 300$ .

**(RRCS):** To quantify the “ripple effect” [1, 40], we evaluate performance on the semantically entangled concepts identified above. A high RRA indicates surgical unlearning with minimal collateral damage. The **Related Retention CLIP Score (RRCS)** further ensures that the visual quality and text-alignment for these related concepts remain degraded.

- **General Retention Accuracy (GRA) & CLIP Score (GRCS):** GRA assesses the model’s stability on a broad set of unrelated concepts to detect catastrophic forgetting. The **General Retention CLIP Score (GRCS)** verifies that the model’s general instruction-following capabilities remain intact after sequential updates.

**Generative Quality.** Beyond accuracy, we follow standard practice to evaluate the fundamental quality of the generated images. We adopt the **Fréchet Inception Distance (FID)** metric, following the specific evaluation script and methodology used in the UnlearnCanvas benchmark [40]. To measure the preservation of distribution quality, we compute the FID between the images generated by our *Unlearned Model* and the images generated by the original *Stable Diffusion* model (acting as the ground truth distribution) for all evaluation prompts. For all CLIP Score calculations (UCS, RRCS, GRCS), we utilize the **openai/clip-vit-large-patch14** model to ensure alignment with standard evaluation protocols.

## 10. Ablation Study on Timestep Range

The choice of the timestep range  $t$  during the distillation process is a critical hyperparameter. It governs the trade-off between the model’s plasticity (ability to unlearn) and stability (ability to retain). To determine the optimal range, we conducted a controlled continual unlearning experiment on a shorter sequence of 4 diverse concepts: *Pikachu*, *Dog*, *Brad Pitt*, and *Van Gogh Style*.

We compared four settings for the maximum sampling timestep  $T_{\max}$ : 300, 600, 800, and 1000. The results, averaged across all concepts after the sequential process, are summarized in Table 4.

### Analysis of Trade-offs:

- **Low Timesteps ( $T = 300$ ):** Restricting training to the later denoising steps (low noise) results in excellent reten-

tion (RRA 0.88), as the global semantic structure formed at higher noise levels remains untouched. However, this comes at the cost of poor unlearning efficacy (0.82), as the model fails to erase the high-level semantic concepts effectively.

- **High Timesteps ( $T = 800$ ):** Increasing the range to include higher noise levels significantly boosts unlearning efficacy (0.93), as it allows the model to modify the concept formation at its source. However, we observe a slight drop in related concept retention (RRA 0.85), indicating a larger “ripple effect”.
- **Convergence Issues ( $T = 1000$ ):** Training on the full range ( $T = 1000$ ) proved unstable in a continual setting. As seen in the table, the unlearning efficacy drops back to 0.82. We observed that the model struggled to converge within the fixed training budget when trying to distill pure noise, leading to inconsistent erasure (e.g., effective on some concepts but failing on others like *Dog* or *Van Gogh*).
- **Optimal Selection ( $T = 600$ ):** We selected  $T = 600$  for all main experiments. It provides a robust “sweet spot”, achieving high unlearning efficacy (0.91) comparable to aggressive settings, while maintaining stability and retention scores (RRA 0.86) close to the conservative settings.

## 11. Main Quantitative Results

We present the detailed breakdown of metrics for sequential unlearning tasks in Table 5.

## 12. Additional Qualitative Results

We provide additional qualitative comparisons in Figures 4–13. Here we visualize the outputs from across the Continual Unlearning Process. Fig 4 shows images generated by the model after unlearning just Pikachu. Similarly, Fig 5 shows images generated by the model after unlearning both Pikachu and Brad Pitt in a sequential manner. Fig 13 shows the images generated by the model after unlearning all 10 concepts sequentially in the following order: *Pikachu*, *Brad Pitt*, *Dog*, *Golf Ball*, *Van Gogh Style*, *Apple*, *Spiderman*, *Lionel Messi*, *Cartoon Style*, and *Banana*.

Method	Concept	Accuracy ↑			CLIP Score ↑			FID ↓
		UA	RRA	GRA	UCS	RRCS	GRCS	
DUGE	Pikachu	0.96	0.65	0.83	29.9	33.4	32.5	6.4
	Brad Pitt	0.82	0.77	0.83	31.8	34.2	32.4	6.3
	Dog	0.54	0.84	0.81	31.5	34.1	32.3	6.5
	Golf Ball	0.60	0.74	0.81	31.7	33.2	32.4	7.5
	Van Gogh Style	0.48	0.79	0.80	32.7	33.3	32.2	7.3
	Apple	0.58	0.75	0.78	32.1	32.8	32.3	8.8
	Spiderman	0.64	0.73	0.74	31.6	32.9	31.9	10.0
	Lionel Messi	0.66	0.70	0.75	31.5	32.8	31.8	10.3
	Cartoon Style	0.59	0.72	0.74	31.6	32.7	31.8	10.4
	Banana	0.63	0.70	0.77	31.1	32.5	31.8	11.2
ESD-u	Pikachu	0.97	0.54	0.75	29.3	31.6	31.8	11.46
	Brad Pitt	0.77	0.59	0.67	30.1	31.1	30.6	21.8
	Dog	0.96	0.03	0.08	19.1	18.6	19.9	264.9
	Golf Ball	0.67	0.38	0.45	27.1	26.8	26.8	82.5
	Van Gogh Style	0.53	0.49	0.54	29.1	28.9	28.5	48.0
	Apple	0.67	0.33	0.32	25.6	24.5	24.7	129.6
	Spiderman	0.73	0.24	0.21	24.3	23.1	23.1	156.0
	Lionel Messi	0.79	0.18	0.17	23.3	21.9	22.3	180.9
	Cartoon Style	0.83	0.14	0.13	22.2	20.9	21.6	212.8
	Banana	0.87	0.12	0.12	21.5	20.6	21.4	222.1
ESD-x	Pikachu	0.99	0.51	0.77	28.1	31.1	31.8	7.7
	Brad Pitt	0.92	0.58	0.70	28.0	31.5	30.9	10.3
	Dog	0.92	0.20	0.18	21.2	22.1	22.3	53.9
	Golf Ball	0.78	0.33	0.41	25.5	27.1	27.3	32.1
	Van Gogh Style	0.71	0.51	0.56	27.9	29.8	29.3	18.0
	Apple	0.83	0.38	0.39	25.4	26.9	27.1	33.6
	Spiderman	0.86	0.33	0.35	24.6	26.2	26.0	39.0
	Lionel Messi	0.83	0.34	0.36	25.0	26.2	26.0	38.3
	Cartoon Style	0.86	0.28	0.29	23.8	24.9	25.1	46.0
	Banana	0.87	0.26	0.26	23.0	24.2	24.4	49.8
MACE	Pikachu	0.99	0.61	0.81	27.6	32.7	32.1	6.8
	Brad Pitt	0.53	0.86	0.86	31.6	34.5	32.2	6.1
	Dog	0.43	0.86	0.82	30.5	33.6	31.9	6.8
	Golf Ball	0.61	0.65	0.74	29.5	31.2	31.1	10.1
	Van Gogh Style	0.22	0.84	0.84	33.2	33.3	32.1	6.2
	Apple	0.41	0.78	0.74	31.4	31.5	31.1	8.6
	Spiderman	0.28	0.85	0.80	32.5	32.9	31.8	6.6
	Lionel Messi	0.21	0.81	0.83	32.2	32.8	31.8	6.7
	Cartoon Style	0.15	0.85	0.83	32.9	33.0	31.9	6.0
	Banana	0.29	0.81	0.77	31.8	32.1	31.4	7.7
UCE	Pikachu	1.00	0.65	0.36	21.2	25.9	25.0	52.5
	Brad Pitt	1.00	0.34	0.34	20.4	22.9	24.9	55.4
	Dog	1.00	0.12	0.09	18.5	19.7	21.1	137.2
	Golf Ball	1.00	0.07	0.07	18.7	19.9	20.6	158.0
	Van Gogh Style	0.89	0.13	0.10	19.4	20.4	21.2	111.5
	Apple	0.92	0.09	0.04	19.4	20.0	20.2	132.9
	Spiderman	0.94	0.07	0.04	19.1	19.6	20.4	137.0
	Lionel Messi	0.99	0.06	0.04	18.6	19.3	20.6	136.8
	Cartoon Style	0.98	0.07	0.02	18.9	19.7	20.7	130.0
	Banana	0.98	0.05	0.02	18.9	19.8	20.9	146.7
Our Method (Fixed Context Mapping)	Pikachu	0.85	0.88	0.88	33.2	35.3	32.6	8.00
	Brad Pitt	0.85	0.86	0.87	31.3	34.9	32.4	8.0
	Dog	0.92	0.88	0.87	30.2	34.4	32.4	8.4
	Golf Ball	0.93	0.85	0.87	30.5	33.6	32.5	8.8
	Van Gogh Style	0.85	0.86	0.86	30.8	33.5	32.3	9.2
	Apple	0.86	0.87	0.85	30.8	33.1	32.4	9.4
	Spiderman	0.86	0.86	0.83	30.7	33.2	32.2	9.7
	Lionel Messi	0.89	0.81	0.84	30.4	33.1	32.1	9.9
	Cartoon Style	0.85	0.82	0.84	30.5	33.0	32.1	10.1
	Banana	0.86	0.81	0.85	30.4	33.0	32.1	10.5
Our Method (Adaptive Context Mapping)	Pikachu	0.97	0.86	0.87	30.2	35.2	32.5	7.9
	Brad Pitt	0.97	0.86	0.86	29.1	34.8	32.3	8.0
	Dog	0.92	0.89	0.84	28.6	34.4	32.2	8.2
	Golf Ball	0.93	0.84	0.85	30.4	33.3	32.4	8.6
	Van Gogh Style	0.83	0.85	0.83	29.6	33.5	32.2	9.3
	Apple	0.88	0.84	0.84	30.7	32.9	32.2	9.6
	Spiderman	0.86	0.85	0.82	30.6	33.1	32.1	10.0
	Lionel Messi	0.90	0.79	0.82	30.3	33.1	32.4	10.3
	Cartoon Style	0.81	0.78	0.81	29.5	32.7	31.7	10.7
	Banana	0.86	0.80	0.83	30.2	32.8	32.0	11.1

Table 5. Comparison of different methods across 10 concepts in continual manner. All metrics should be high except FID (lower is better).

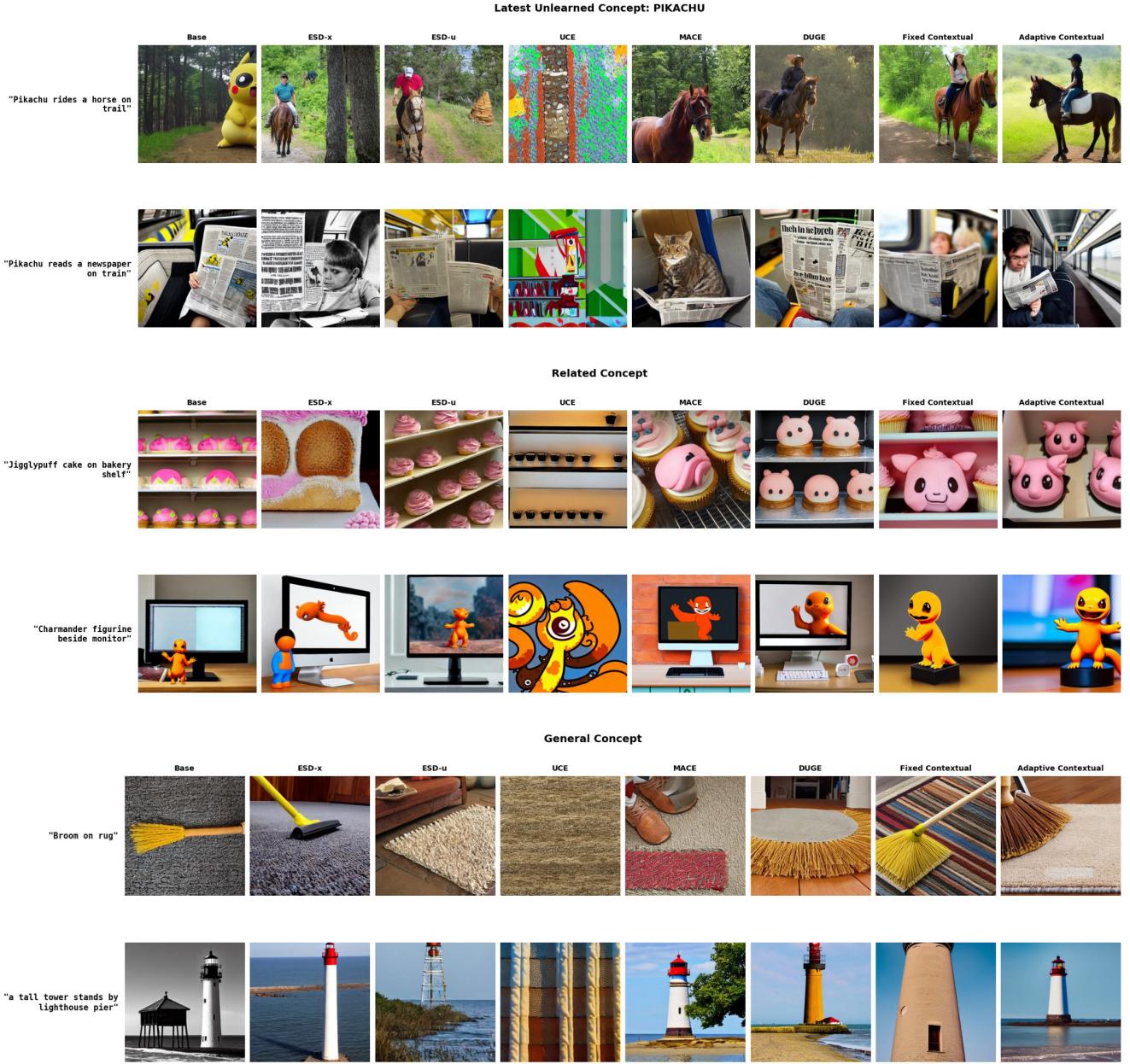


Figure 4. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

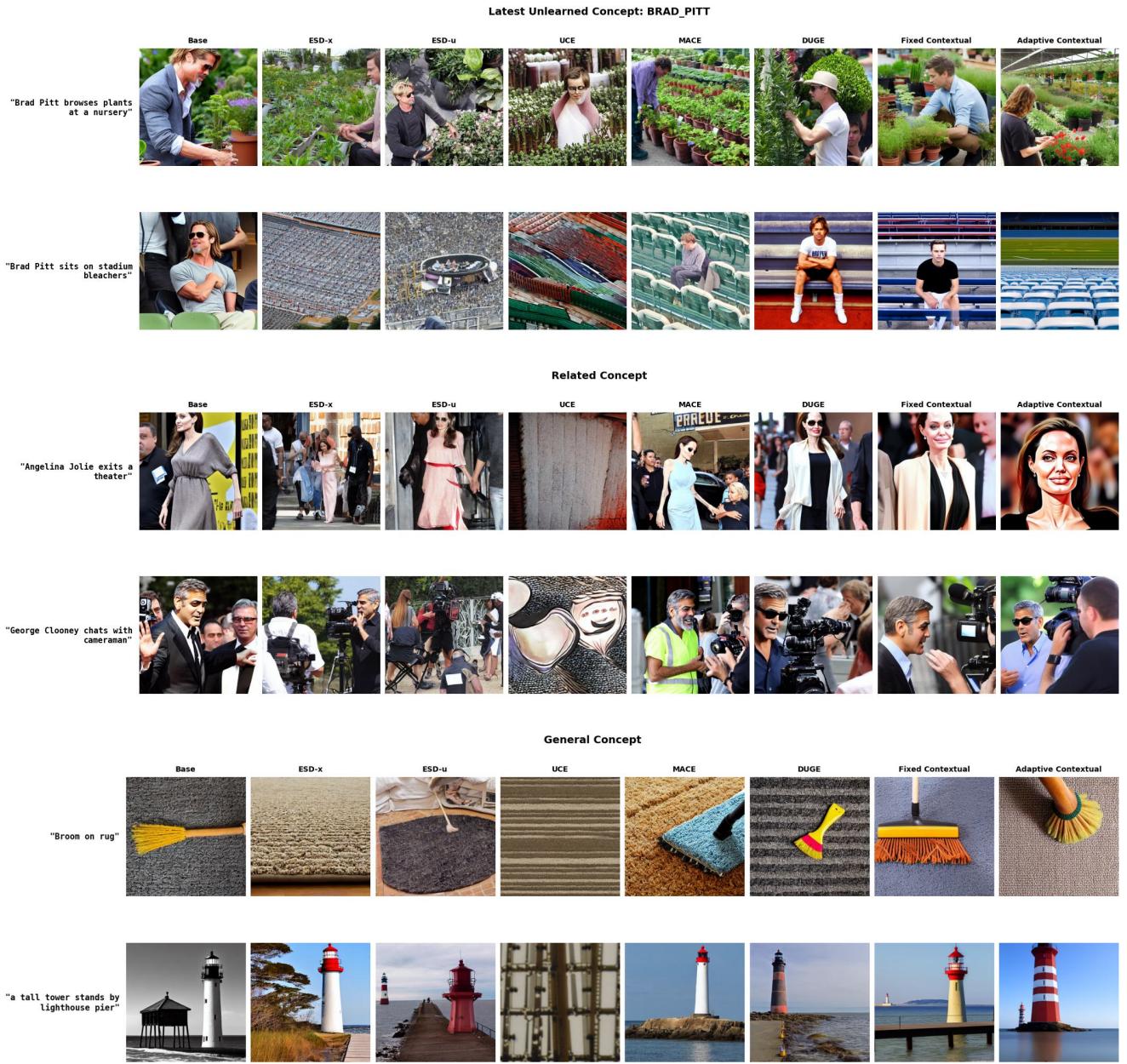


Figure 5. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

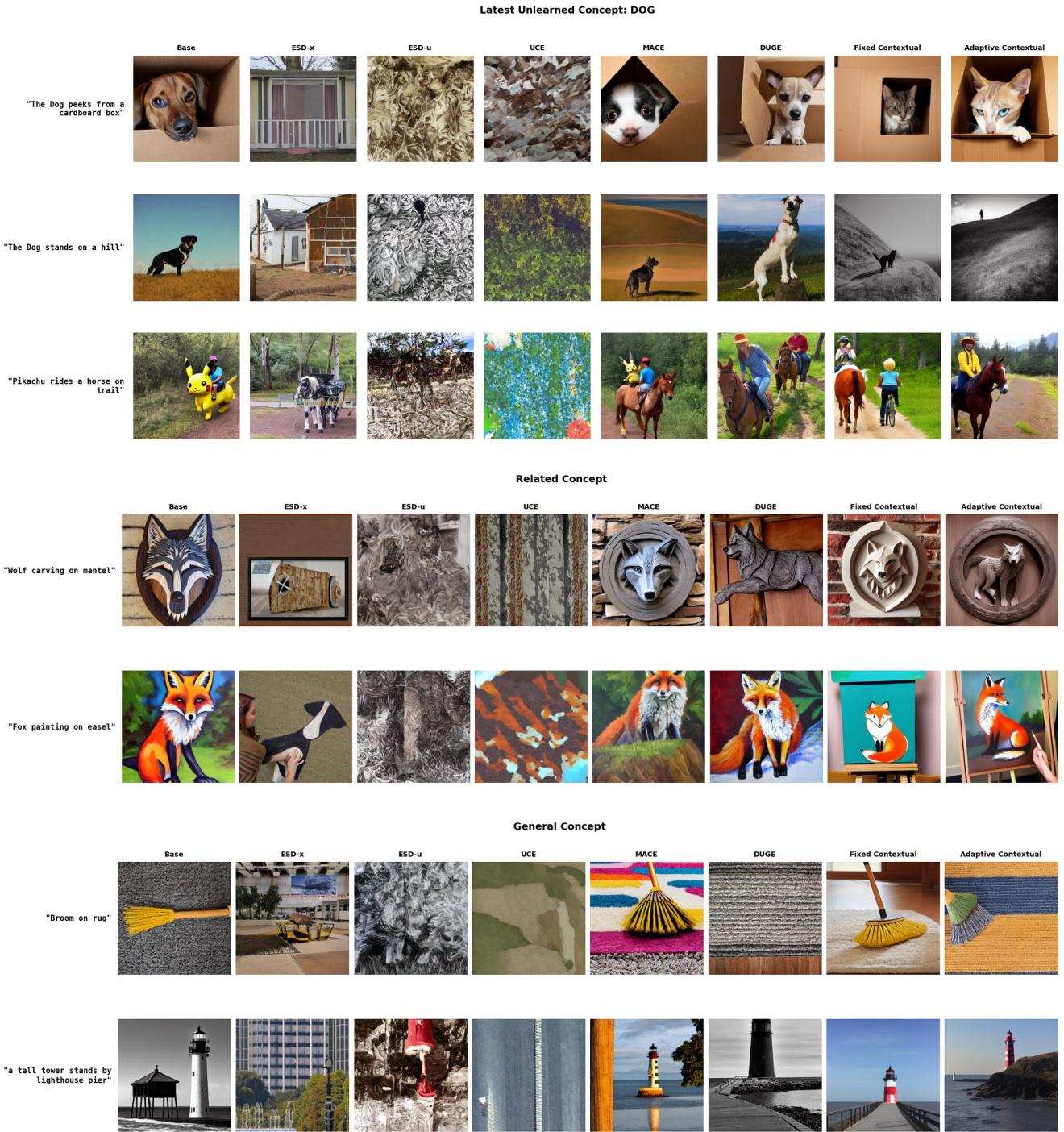
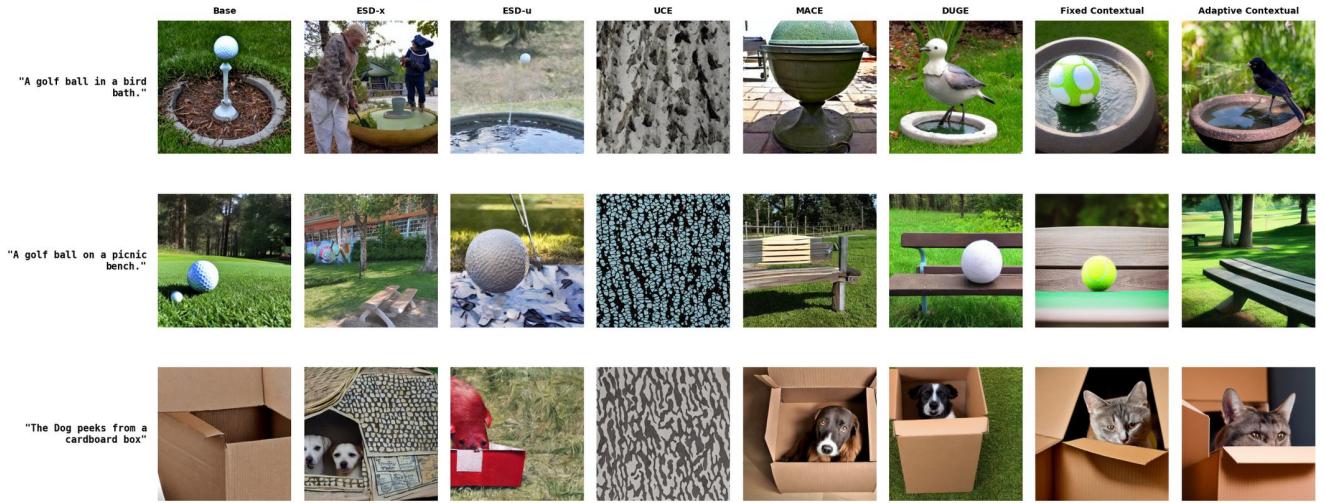
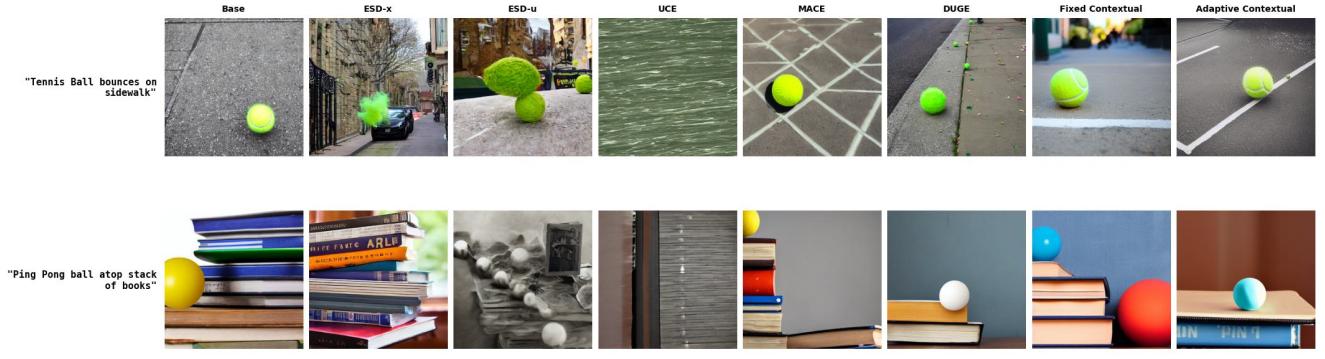


Figure 6. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

#### Latest Unlearned Concept: GOLF\_BALL



#### Related Concept



#### General Concept



Figure 7. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

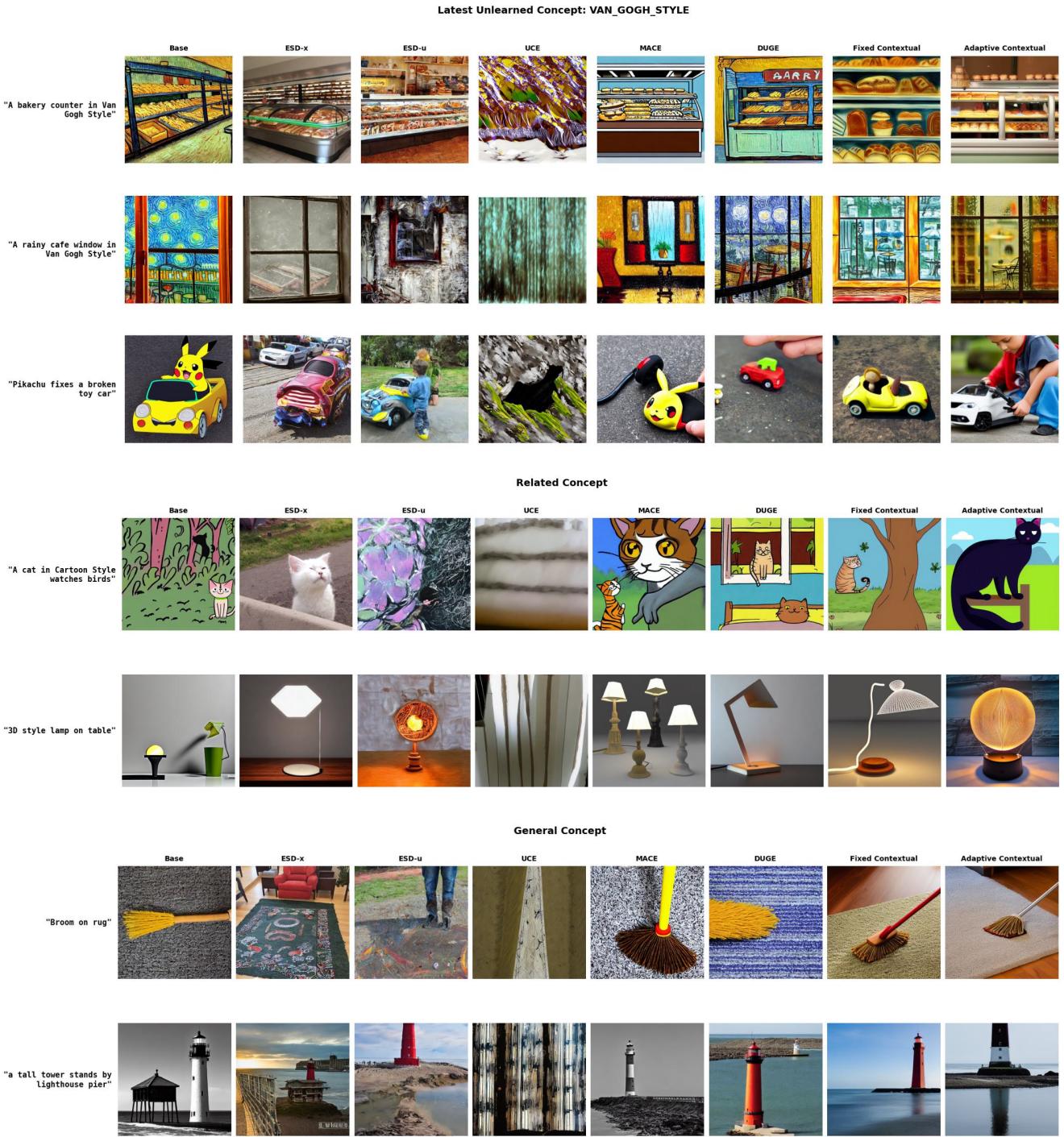


Figure 8. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

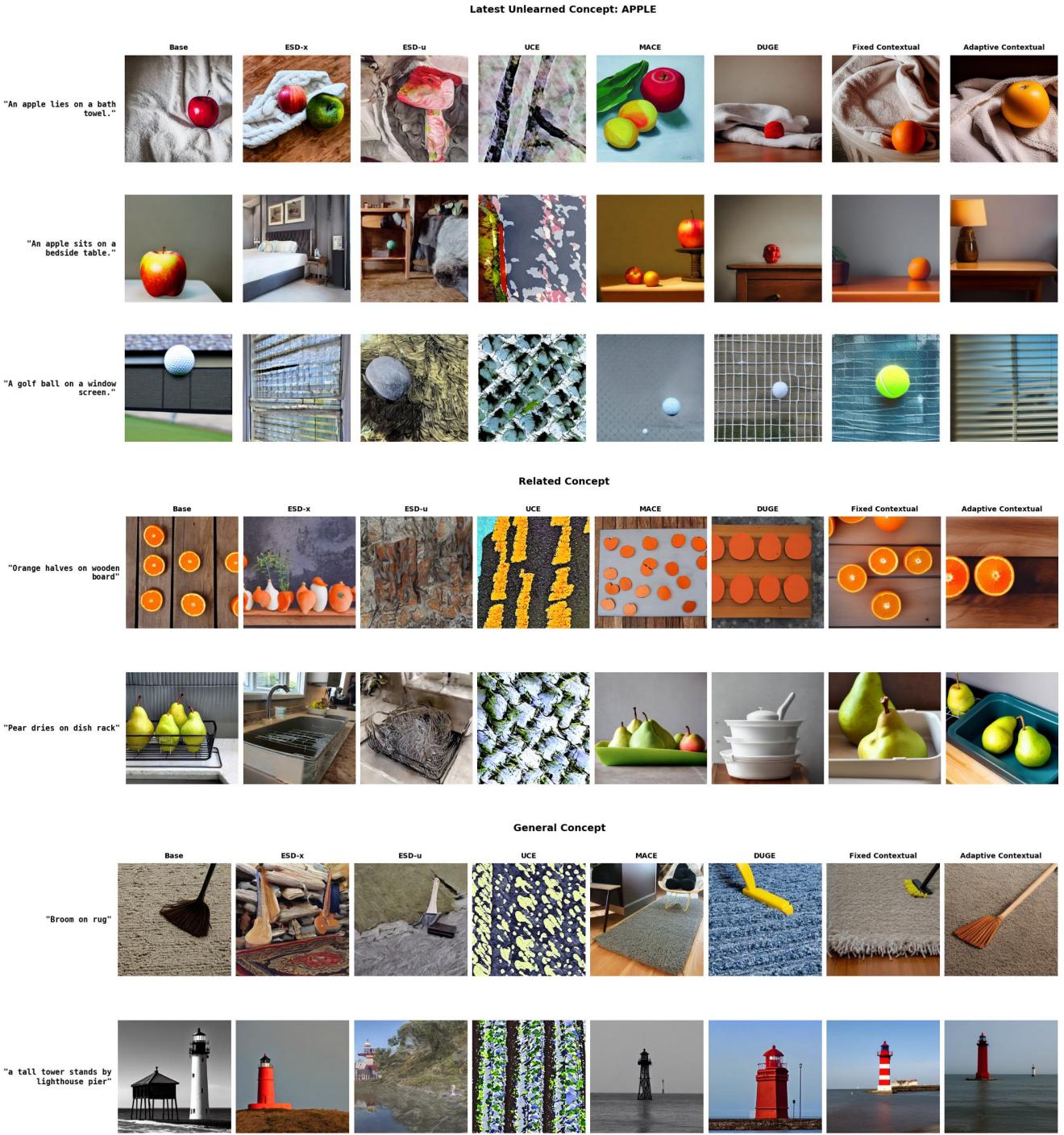


Figure 9. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

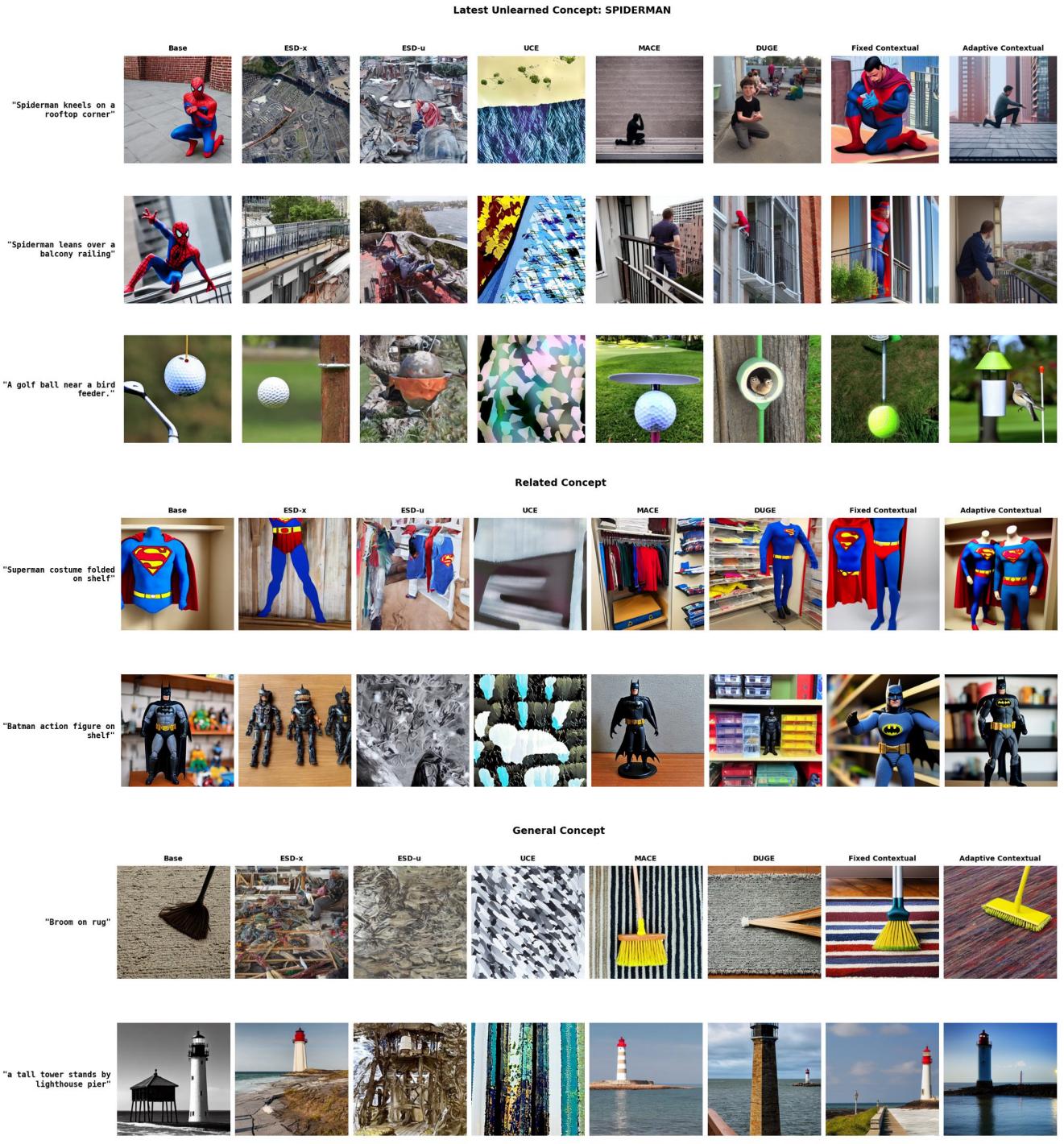


Figure 10. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

#### Latest Unlearned Concept: LIONEL\_MESSI

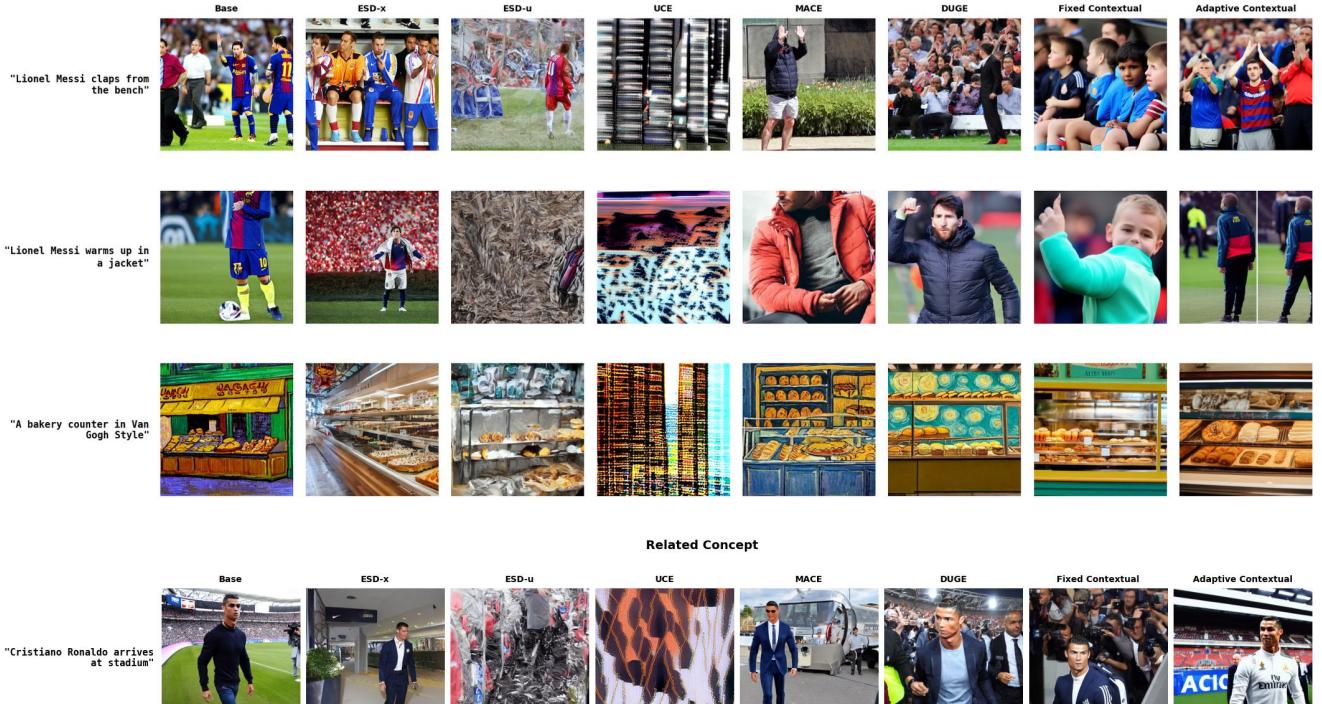


Figure 11. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

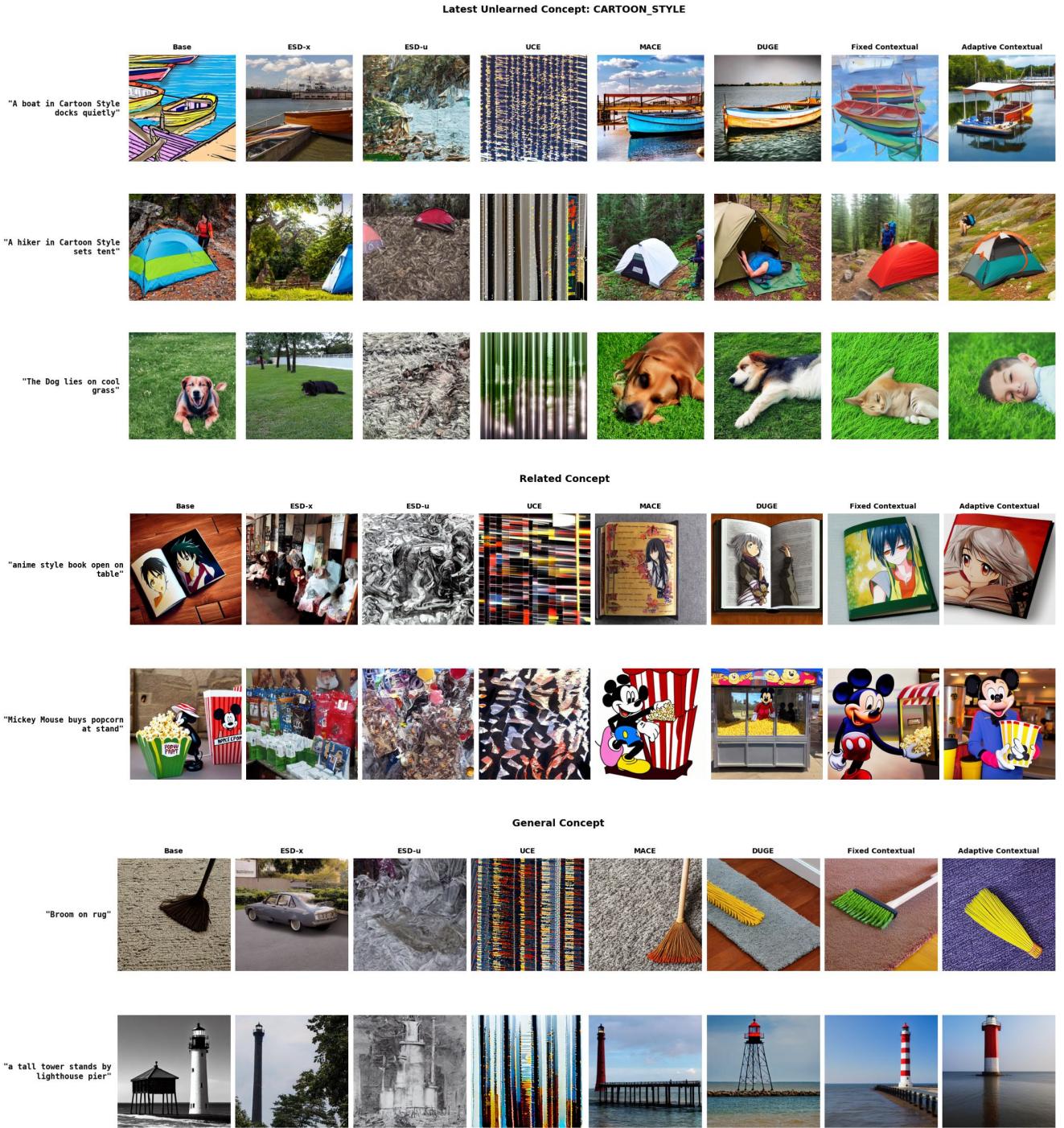


Figure 12. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.

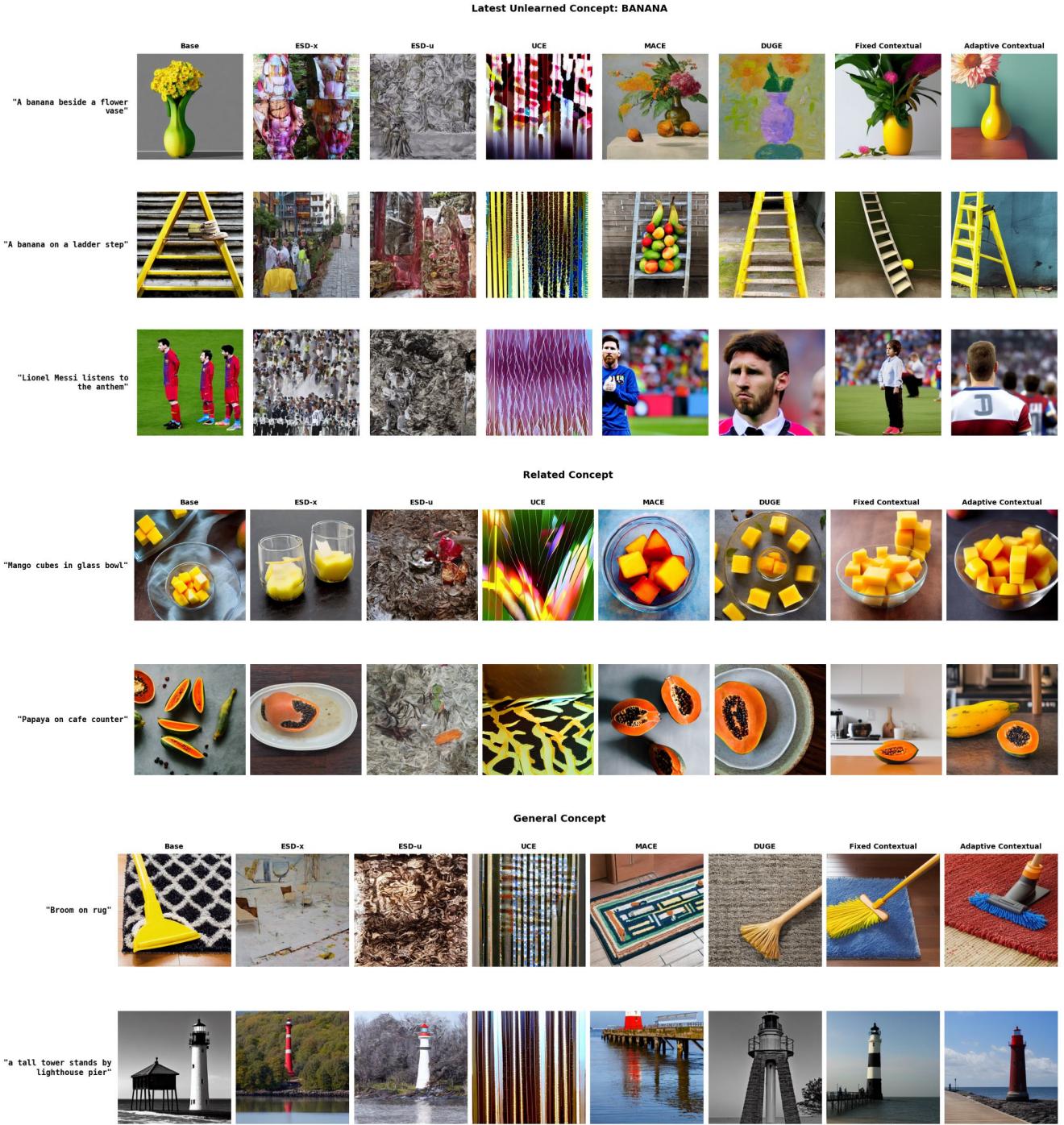


Figure 13. Visualization comparing the effects of unlearning across multiple SOTA methods on both the unlearned and retain sets. Rows show different prompts, columns show different methods: *Base* (Stable Diffusion v1.5), ESD-x, ESD-u, UCE, MACE, DUGE, Fixed Contextual, and Adaptive Contextual. Three sections display: unlearned concept images, related concept images, and general retain set images, demonstrating how each method balances effective unlearning with preservation of model capabilities.