FILTERING MEMORIZATION FROM PARAMETER-SPACE IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-Rank Adaptation (LoRA) enables efficient customization of diffusion models, but may also memorize and reproduce sensitive or copyrighted training content. This risk is particularly relevant in LoRA-sharing ecosystems, where adapters are distributed without their training data or training pipelines. Existing mitigation methods typically require access to training data, intervention during training, or control over inference, making them unsuitable when only a public pretrained backbone and a third-party LoRA are available. We propose **Base-Anchored Filtering (BAF)**, a training- and data-free framework for post-hoc memorization mitigation. BAF spectrally decomposes the LoRA update and measures each component's alignment with the pretrained backbone's principal subspace, preserving strongly anchored components while suppressing weakly anchored components associated with instance-specific memorization. Experiments across three datasets, three diffusion backbones, and 20 community-released LoRAs show that BAF consistently reduces memorization while preserving generation quality, demonstrating that pretrained parameter geometry provides a useful structural prior for filtering memorization from released LoRAs. Our code is in the supplementary material.

1 INTRODUCTION

Text-to-image diffusion models (Ho et al., 2020; Rombach et al., 2022) have become the dominant paradigm for high-fidelity image generation. Building on pretrained backbones such as Stable Diffusion (Rombach et al., 2022), users increasingly customize generation toward specific styles, characters, or visual concepts. Low-Rank Adaptation (LoRA) (Hu et al., 2021) enables such customization by fine-tuning a small set of low-rank weight updates from only a few reference images. This simplicity has given rise to a large-scale open ecosystem. Platforms such as Hugging Face and CivitAI host hundreds of thousands of user-uploaded LoRAs[1], many of which are publicly available, with direct plug-and-play usage on popular diffusion backbones.

However, LoRAs may memorize and reproduce training examples or near-identical variants (Carlini et al., 2023; Somepalli et al., 2023a; Wen et al., 2024). When such content includes copyrighted artwork, protected characters, or identifiable individuals, this may lead to unintended reproduction. The risk is particularly relevant in LoRA-sharing scenarios, where downstream users and hosting platforms have access only to the released adapter weights, without access to the training data or optimization process or prior knowledge of what content has been memorized, while such content may still surface during generation (Fig. 1).

Although several methods have been proposed to mitigate memorization, they are poorly suited to the LoRA-sharing scenarios. Training-time defenses (Somepalli et al., 2023a; Tsai et al., 2025) mitigate memorization by modifying the training data or optimization process, but are unavailable once only the final LoRA weights are released. Output-based auditing methods detect memorization by comparing generated samples against known training images (Carlini et al., 2023; Wen et al., 2024), while concept-erasure methods remove specified learned concepts directly from model weights (Gandikota et al., 2023; 2024). Both therefore rely on prior knowledge of the content that may have been memorized, which is unavailable when a third-party LoRA is uploaded without its training data. Inference-time defenses instead modify the sampling process to suppress memorized

[1] CivitAI: https://civitai.com; Hugging Face: https://huggingface.co/

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

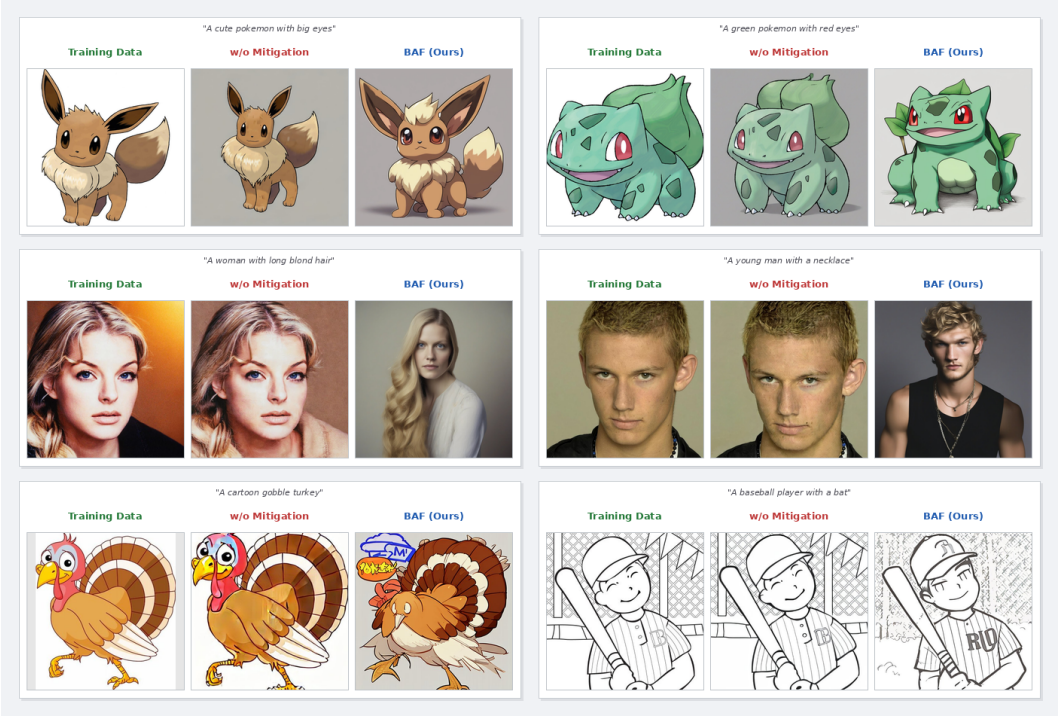

Figure 1: LoRA weight carries memorization of training data. BAF successfully modifies diffusion models to produce memorization-free outputs.

outputs (Chen et al., 2024b). Although such defenses can be adopted by users who control the generation pipeline, they cannot be reliably enforced by hosting platforms whose role is limited to receiving and distributing LoRA weights. We therefore ask: *Can memorization be identified and selectively removed using only the released LoRA parameters?*

Our key observation is that the pretrained backbone provides a useful structural anchor. Large-scale pretraining induces organized spectral structure in model parameters (Kwon et al., 2022; Si et al., 2024), while meaningful task-specific variation often lies in low-dimensional subspaces aligned with pretrained directions (Aghajanyan et al., 2021). This motivates the intuition that *generalizable adaptation tends to follow pretrained spectral directions, whereas instance-specific memorization is more weakly aligned*. Empirically, we observe that spectral channels with high alignment preserve the intended adaptation with little memorization, while weakly aligned channels retain substantially more memorized content.

Building on this observation, we propose **Base-Anchored Filtering (BAF)**, a training- and data-free framework for post-hoc memorization mitigation in diffusion LoRAs. BAF decomposes a LoRA update into spectral channels, measures their alignment with the pretrained model's principal subspace, and suppresses weakly anchored channels directly in parameter space. It requires neither training data, memorized examples, nor control over downstream inference.

Our contributions are as follows:

1. We uncover a geometric property showing that spectral alignment with the pretrained principal subspace separates generalizable adaptation from memorization.
2. Based on this observation, we propose *Base-Anchored Filtering* (BAF), a training- and data-free framework that removes memorization directly from released LoRA weights in the realistic weight-only setting.
3. We provide a theoretical analysis showing that BAF succeeds when memorization and benign adaptation are sufficiently separable in parameter space, and empirically verify that this condition holds in one controlled setting.

4. We conduct extensive experiments on controlled memorization benchmarks and community-released LoRAs across multiple diffusion backbones, demonstrating that BAF consistently achieves a superior memorization–utility trade-off.

2 Related Work

2.1 Memorization in diffusion model

Memorization in diffusion models refers to the tendency of a model to reproduce training images or near-identical variants during generation. Early studies (Somepalli et al., 2023a; Carlini et al., 2023) showed that text-to-image systems such as Stable Diffusion replicate a non-trivial portion of their training data. Subsequent work investigated the mechanisms behind memorization. (Gu et al., 2023) analyzed the effects of dataset size, duplication rate, and model capacity, and observed that in LoRA fine-tuning, memorization strength increases with the rank of the adaptation matrices. At a different granularity, (Wen et al., 2024) linked memorization to abnormal text-conditional prediction magnitudes and proposed token-level attribution to identify triggering words. (Ren et al., 2024) localized memorization effects to cross-attention patterns, while (Chen et al., 2024a) distinguished between *global* and *local* memorization, showing that the latter is harder to detect and mitigate. From a geometric perspective, (Ross et al., 2024) connected memorization to regions of near-zero Local Intrinsic Dimensionality (LID).

2.2 Memorization Detection and Mitigation

Existing memorization mitigation methods can be broadly categorized into training-time, inference-time, and parameter-space approaches.

Training-time approaches mitigate memorization by modifying the training data or optimization process. Data deduplication (OpenAI, 2022; Kandpal et al., 2022) removes near-duplicate samples, while differential privacy (Tsai et al., 2025; Dockhorn et al., 2022) limits the influence of individual examples. These methods require access to training and are therefore unsuitable when only the final LoRA weights are available.

Inference-time approaches suppress memorized outputs during sampling. RTA (Somepalli et al., 2023b) perturbs prompts, prompt optimization (Wen et al., 2024) weakens memorization-triggering conditioning signals, CFG-based methods (Jain et al., 2025) modify guidance, and AMG (Chen et al., 2024b) detects and corrects replication during generation. However, these methods require control over inference and leave memorized information intact in the model parameters.

Parameter-space approaches directly edit model weights. Concept-erasure methods (Gandikota et al., 2023; 2024) remove specified concepts, while neuron- or weight-level editing methods (Hintersdorf et al., 2024; Chavhan et al., 2024) identify and suppress components associated with memorized prompts. These methods still require prior knowledge of the memorized concepts, prompts, or examples. In contrast, our setting assumes access only to the released LoRA weights and the corresponding pretrained backbone.

3 Preliminaries

Low-Rank Adaptation. LoRA (Hu et al., 2021) fine-tunes a pretrained model by adding trainable low-rank updates to selected layers. For a pretrained weight matrix $\boldsymbol{W}_* \in \mathbb{R}^{m\times n}$, LoRA introduces $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$ with rank $r \ll \min(m, n)$, and the adapted weight is $\boldsymbol{W} = \boldsymbol{W}_* + \Delta\boldsymbol{W}$ where $\Delta\boldsymbol{W} \triangleq BA$ is the additive weight update. To analyze the internal structure of LoRA updates, we turn to their spectral decomposition.

Spectral decomposition of LoRA updates. The singular value decomposition (SVD) of a rank-r LoRA update is $\Delta\boldsymbol{W} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are singular values and $\boldsymbol{u}_i \in \mathbb{R}^m$, $\boldsymbol{v}_i \in \mathbb{R}^n$ are left/right singular vectors. Each rank-1 term $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ is a *spectral channel*: it maps an input direction $\boldsymbol{v}_i$ to an output direction $\boldsymbol{u}_i$ with magnitude σ_i.

Top-K principal subspace of the pretrained model. Let $\boldsymbol{W}_* \in \mathbb{R}^{m\times n}$ denote the pretrained weight at the same layer. We compute its singular value decomposition (SVD):

$$\boldsymbol{W}_* = \boldsymbol{U}_*\boldsymbol{\Sigma}_*\boldsymbol{V}_*^\top = \sum_{j=1}^{\min(m,n)} s_j\, \boldsymbol{u}_j^*\boldsymbol{v}_j^{*\top},$$

where $\boldsymbol{U}_* = [\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_m^*]$ and $\boldsymbol{V}_* = [\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_n^*]$ are orthonormal matrices and $\boldsymbol{\Sigma}_* = \mathrm{diag}(s_1, \ldots, s_{\min(m,n)})$ contains the singular values.

Let $\boldsymbol{U}_{*,K}$ and $\boldsymbol{V}_{*,K}$ denote the matrices containing the first K left and right singular vectors of $\boldsymbol{W}_*$. We define the orthogonal projection operators onto the corresponding principal subspaces as

$$\boldsymbol{P}_L = \boldsymbol{U}_{*,K}\boldsymbol{U}_{*,K}^\top, \qquad \boldsymbol{P}_R = \boldsymbol{V}_{*,K}\boldsymbol{V}_{*,K}^\top.$$

Problem Setup. We consider a setting where only a released LoRA adapter is available. The adapter is associated with a publicly available pretrained base model with weights $\boldsymbol{W}_*$. For each adapted layer, the LoRA update ΔW is observable, while the fine-tuning data, optimization procedure, and hyperparameters used to obtain it are unavailable. Moreover, there is no prior knowledge of whether the adapter memorizes any training examples or what those examples may contain.

This setting reflects current LoRA-sharing ecosystems such as CivitAI, where user-trained adapters are distributed independently of their training data. Consequently, both hosting platforms and downstream users may need to inspect or modify an adapter using only the released LoRA weights together with the corresponding public pretrained model.

Given $\boldsymbol{W}_*$ and ΔW, our objective is to construct a filtered update $\Delta W_{\mathrm{filtered}}$ that preserves the generalizable adaptation encoded by the LoRA (e.g., artistic styles or visual concepts) while suppressing instance-specific memorization (e.g., near-verbatim reproductions of copyrighted training images). The filtering procedure must therefore distinguish reusable adaptation from memorization without access to training data, reference images, or any additional optimization.

4 Main Method

We propose **BAF** (Base-Anchored Filtering), a data-free framework for mitigating memorization in LoRA weights using only parameter-space structure. Our key insight is that spectral channels aligned with the pretrained model's principal subspace tend to encode *generalizable adaptation*, while weakly aligned channels are more likely to capture *memorized content*.

In this section, we start off by defining an *anchoring score* to measure spectral alignment and empirically validate the insight via a directional ablation experiment. We then present the complete BAF filtering algorithm. Then, to improve on the performance of **BAF**, we introduce an *adaptive strategy* for selecting the principal-subspace dimension K across different architectures, based on an energy-adaptive rule. Finally, we provide a theoretical analysis of why BAF works.

4.1 Empirical Motivation

To validate our insight directly, we first define a quantitative measure of spectral alignment.

Definition 1 (Anchoring score)**.** For the i-th spectral channel with left and right singular vectors $\boldsymbol{u}_i \in \mathbb{R}^m$ and $\boldsymbol{v}_i \in \mathbb{R}^n$, the *anchoring score* with respect to the pretrained model $\boldsymbol{W}_*$'s top-K principal subspace is defined as

$$a_i = \|\boldsymbol{P}_L\boldsymbol{u}_i\|^2 \cdot \|\boldsymbol{P}_R\boldsymbol{v}_i\|^2 \in [0,1], \tag{1}$$

where $\boldsymbol{P}_L$ and $\boldsymbol{P}_R$ are the orthogonal projections onto the pretrained left and right principal subspaces.

Remark 1. The anchoring score admits a simple geometric interpretation. Since singular vectors are unit vectors, the projection norm satisfies $\|\boldsymbol{P}_L\boldsymbol{u}_i\|^2 = \cos^2\theta_i^L$, where θ_i^L is the angle between $\boldsymbol{u}_i$ and the pretrained left subspace. Thus $a_i = \cos^2\theta_i^L \cdot \cos^2\theta_i^R \in [0,1]$ measures the bilateral alignment of the i-th spectral channel with the pretrained subspaces. Channels with large a_i lie largely within the pretrained directions, whereas channels with small a_i introduce directions that are largely orthogonal to the pretrained model.

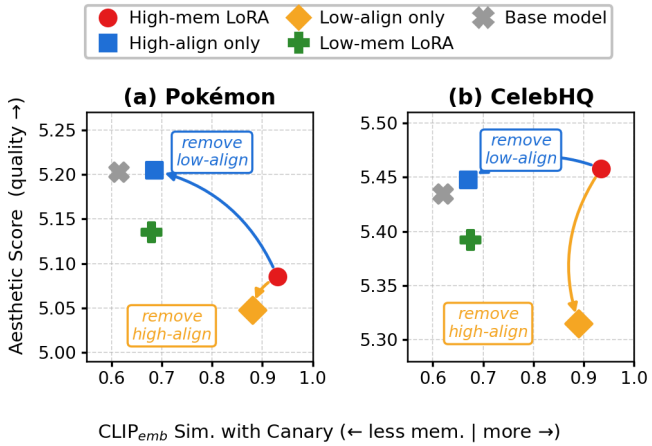

Figure 2: Retaining only highly aligned channels suppresses memorization while preserving generation quality, whereas retaining only weakly aligned channels retains memorization but substantially degrades generation quality. Results are consistent across Pokémon (a) and CelebA-HQ (b).

This geometric view is consistent with recent analyses of geometric memorization (Ross et al., 2024), which associate memorized samples with low Local Intrinsic Dimensionality (LID). Such instance-specific information is less likely to align with the pretrained principal subspace, resulting in low anchoring scores. The anchoring score can therefore be viewed as a parameter-space analogue of LID, providing a complementary perspective on memorization.

Definition 2 (Random-projection baseline)**.** If $\boldsymbol{u}$ and $\boldsymbol{v}$ are independently drawn uniformly from the unit spheres in $\mathbb{R}^m$ and $\mathbb{R}^n$, then the expected anchoring score is

$$a_{\text{null}} \triangleq \mathbb{E}[a(\boldsymbol{u}, \boldsymbol{v})] = \frac{K}{m} \cdot \frac{K}{n} = \frac{K^2}{mn}. \tag{2}$$

Remark 2. The baseline a_{null} represents the expected anchoring score of a channel that carries no structured relationship to the pretrained model. Intuitively, if a spectral channel does not reuse the pretrained directions, its singular vectors behave like arbitrary directions in weight space. Modeling them as random unit vectors provides a natural reference point for alignment. Channels with $a_i \gg a_{\text{null}}$ are strongly anchored to the pretrained subspace, whereas channels with scores close to a_{null} exhibit no more alignment than random directions.

Directional ablation experiment. To test whether anchoring scores separate generalizable adaptation from memorization, we follow (Wen et al., 2024) and train a *High-mem LoRA* by duplicating 50 canary images 50 times, together with a *Low-mem LoRA* without duplication. We partition the High-mem LoRA's spectral channels using a_{null} and construct *high-align-only* ($a_i \geq a_{\text{null}}$) and *low-align-only* ($a_i < a_{\text{null}}$) variants.

Metrics. We measure memorization using CLIP similarity to the duplicated canaries and generation quality using Aesthetic Score (AES) (LAION-AI, 2022).

Results. Figure 2 shows that high-alignment channels reduce memorization while preserving generation quality, whereas low-alignment channels retain substantially more memorization and degrade quality. This trend is consistent across Pokémon and CelebA-HQ, supporting our hypothesis that strongly anchored channels primarily encode reusable adaptation, while weakly anchored channels carry more memorized content. Experimental details are provided in Appendix A.

4.2 The BAF Algorithm

Motivated by the above observation, BAF applies anchoring-based gating to each layer of a single LoRA update. We define two gating modes.

BAF-Hard. Binary gating with threshold a_{null}, $g_{\text{hard}}(a_i) = \mathbf{1}[a_i \geq a_{\text{null}}]$. BAF-Hard explicitly removes channels below the anchoring threshold.

BAF-Soft. While BAF-Hard may drastically reduce memorization, it could inadvertently remove useful channels near the anchoring threshold. BAF-Soft instead suppresses rather than removes them via continuous gating with exponent $\alpha > 0$, $g_{\text{soft}}(a_i) = a_i^{\alpha}$. Larger α yields stronger suppression; $\alpha = 0$ recovers the unfiltered LoRA update.

Filtered update. BAF produces the filtered LoRA update at each layer:

$$\Delta \boldsymbol{W}_{\text{BAF}} = \sum_{i=1}^{r} \sigma_i \cdot g(a_i) \cdot \boldsymbol{u}_i \boldsymbol{v}_i^{\top}, \tag{3}$$

where the gating function $g(\cdot)$ depends on the method of choice with the final adapted model being $\boldsymbol{W} = \boldsymbol{W}_* + \Delta \boldsymbol{W}_{\text{BAF}}$. The full algorithm is given in Appendix B.

4.3 Energy-Adaptive Principal Subspace Selection

Selecting the principal-subspace dimension K is important for applying BAF across different diffusion backbones. Pretrained architectures exhibit substantially different spectral structures: some layers concentrate most of their energy in only a few dominant singular directions, whereas others distribute their energy across a much larger subspace. Consequently, a globally fixed principal-subspace dimension may be suitable for one architecture but suboptimal for another, requiring architecture-specific hyperparameter tuning.

Energy-adaptive selection rule. To address this issue, we determine K according to the cumulative spectral energy of each pretrained layer:

$$K = \min \left\{ k \in \{1, \ldots, \min(m, n)\} : \frac{\sum_{j=1}^{k} s_j^2}{\sum_{j=1}^{\min(m,n)} s_j^2} \geq \tau_{\text{energy}} \right\}. \tag{4}$$

where s_j are the singular values of $\boldsymbol{W}_*$ and $\tau_{\text{energy}} \in (0, 1)$ is the target energy ratio.

Remark 3. Adaptive K makes BAF architecture-aware rather than architecture-specific. Instead of relying on a manually chosen subspace dimension, the principal subspace is automatically determined by the intrinsic spectral geometry of each pretrained layer. As a result, the same filtering criterion can be applied consistently across different diffusion backbones without architecture-specific tuning. Layers with rapidly decaying spectra therefore yield smaller values of K, whereas layers with flatter spectra naturally retain larger principal subspaces.

4.4 Theoretical Discussion

To understand why BAF can suppress memorization while preserving useful adaptation, we analyze it under the controlled memorization setting widely used in prior work (Somepalli et al., 2023a; Wen et al., 2024), where a small set of canary images is intentionally duplicated during LoRA fine-tuning. Following this setup, we define the *counterfactual clean LoRA*, denoted by ΔW_{clean}, as the LoRA that would have been obtained from the corresponding training set without duplicated canaries. The observed LoRA can then be conceptually written as

$$\Delta W = \Delta W_{\text{clean}} + \Delta W_{\text{mem}},$$

where ΔW_{mem} denotes the additional parameter change associated with duplication-induced memorization. This decomposition is introduced solely as an *analysis framework*: it does not assume that the fine-tuning process itself decomposes linearly. Importantly, ΔW_{clean} and ΔW_{mem} are not observable in the intended weight-only deployment setting and are not required by BAF. At deployment, BAF operates only on the released LoRA update ΔW and the corresponding pretrained weights. The counterfactual decomposition instead provides a reference for characterizing sufficient conditions under which anchoring-based filtering can recover useful adaptation. The complete setup and derivation are provided in Appendix C.

Our analysis considers the following two structural conditions.

Assumption 1. Generalizable adaptation is predominantly concentrated in strongly anchored spectral channels.

Assumption 2. Memorization is predominantly concentrated in weakly anchored spectral channels.

Together, these assumptions characterize a separation regime in which the anchoring score can distinguish spectral channels associated primarily with memorization from those carrying reusable adaptation.

Theorem 1 (Approximate Recovery)**.** *Under Assumptions 1 and 2, the LoRA produced by BAF satisfies*

$$\|\Delta W_{\mathrm{BAF}} - \Delta W_{\mathrm{clean}}\|_F \leq C_1 \rho_c \|\Delta W_{\mathrm{clean}}\|_F + C_2 \rho_m \|\Delta W_{\mathrm{mem}}\|_F + \kappa, \tag{5}$$

where ρ_c and ρ_m quantify the fractions of benign adaptation and memorization, respectively, that violate the assumed separation, κ captures the irreducible approximation error caused by their intrinsic overlap, and C_1, C_2 depend on the choice of gating function.

The theorem shows that BAF approximately recovers the counterfactual clean LoRA when benign adaptation and memorization are sufficiently separated in anchoring-score space. In Appendix C, we directly examine this condition in a controlled paired setting and observe clear separation between memorization and benign adaptation.

5 Experiments

Scenarios. We evaluate BAF under two complementary scenarios. **Controlled memorization.** For quantitative evaluation, memorization should be both stable and controllable. Following prior work (Wen et al., 2024), we induce memorization by duplicating a set of canary images $\{(x_k, y_k)\}_{k=1}^{N_c}$ during LoRA training, where each canary image is repeated D times. This controlled setup produces LoRAs with predictable memorization strength, enabling quantitative comparisons across different methods. Unless otherwise specified, we use $D = 50$ and $N_c = 50$. **Real-world memorization.** We further evaluate BAF on naturally memorizing LoRAs released by the community, assessing whether its effectiveness generalizes beyond artificially constructed memorization.

Setup. For the controlled memorization scenario, we evaluate on Stable Diffusion v1.5 (Rombach et al., 2022), SDXL (Podell et al., 2023), and Flux.1 (Labs, 2024), using LoRAs applied to the attention projection layers of the corresponding denoising backbone. Experiments are conducted on three datasets: *Pokémon* (Pinkney, 2022), *CelebA-HQ* (Karras et al., 2018), and *LAION Art* (Somepalli et al., 2023b). Detailed training settings are provided in Appendix D. For the real-world scenario, we collect publicly released LoRAs from Civitai spanning multiple generation domains and diffusion backbones, including Stable Diffusion and Flux. Since no training data are available, BAF is applied directly to the released LoRAs, and its ability to suppress memorization while preserving generation quality is evaluated qualitatively.

Evaluation metrics. For the controlled memorization scenario, we use SSCD (Pizzi et al., 2022) to measure memorization, a cosine similarity metric in a copy-detection embedding space. We report **targeted SSCD**, defined as the similarity between each generated image and its corresponding duplicated canary image. Additional memorization metrics (normalized ℓ_2 distance and CLIP embedding distance) are reported in Appendix D. Generation quality is evaluated using CLIP score (Radford et al., 2021) for text-image alignment and aesthetic score (Schuhmann et al., 2022) for visual quality. Additional metrics, including LPIPS (Zhang et al., 2018) and ImageReward (Xu et al., 2023), are provided in Appendix D. For the real-world memorization scenario, where no ground-truth memorized images are available, we conduct a human preference study to evaluate whether BAF suppresses memorized content while preserving the intended generation quality. The detailed evaluation protocol is described in Appendix E.

Baselines. We compare BAF with representative training- and inference-time methods. *Training-time* baselines include **Deduplication** (Somepalli et al., 2023a; Kandpal et al., 2022) and **DP-LoRA** (Tsai et al., 2025); both require access to training and are included only as reference baselines. *Inference-time* baselines include **RTA** (Somepalli et al., 2023b), **Prompt Optim.** (Wen et al., 2024),

Table 1: **Evaluation of memorization mitigation and generation quality when LoRA trained on Pokémon or CelebA-HQ.** Shaded rows refer to methods that are incomparable to our proposal. Best and second-best among comparable (non-shaded) methods are **bolded** and underlined, respectively.

Method	Pokémon									CelebA-HQ								
	SD1.5			SDXL			Flux.1D			SD1.5			SDXL			Flux.1D		
	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑
W/O mitigation	0.76	5.08	33.20	0.74	4.87	33.16	0.81	5.71	34.42	0.93	5.46	29.54	0.84	4.72	29.03	0.88	6.05	32.20
Deduplication	0.65	5.04	33.26	0.63	4.74	32.83	0.70	5.68	34.38	0.53	5.47	29.49	0.56	5.05	28.63	0.59	6.00	32.11
DP-LoRA	0.60	4.85	32.00	0.66	4.40	30.50	0.66	5.20	33.61	0.62	5.25	28.90	0.70	4.60	28.40	0.73	5.58	31.20
RTA	0.72	5.12	32.22	0.68	5.10	30.02	0.77	5.69	34.05	0.74	5.34	29.11	0.81	4.91	28.27	0.84	5.94	31.66
Prompt Optim.	0.66	4.95	31.50	0.71	4.60	29.80	0.72	5.46	33.80	0.68	5.10	28.85	0.74	4.70	27.90	0.77	5.74	31.38
CFG-Reverse	0.79	5.06	33.24	0.78	4.89	33.39	0.82	5.65	34.48	0.71	5.53	29.65	0.75	5.28	30.02	0.79	6.07	32.24
AMG	0.50	5.18	33.05	0.56	5.03	32.82	0.58	5.73	34.33	0.40	5.52	29.72	0.63	5.26	29.84	0.65	6.10	32.18
BAF-Soft	0.57	**5.43**	**33.40**	0.62	**5.71**	**34.27**	0.55	**5.90**	**34.51**	0.52	**5.67**	**31.45**	0.68	**6.13**	**32.45**	0.63	**6.32**	**33.01**
BAF-Hard	**0.54**	5.02	32.38	**0.60**	5.49	34.23	**0.53**	5.73	34.34	**0.44**	5.51	31.24	**0.67**	6.10	32.36	**0.61**	6.17	32.90

Table 2: **Evaluation of memorization mitigation and generation quality when LoRA trained on LAION-Art.** Shaded rows refer to methods that are incomparable to our proposal. Best and second-best among fairly comparable (non-shaded) methods are **bolded** and underlined, respectively.

Method	SD1.5			SDXL			Flux.1D		
	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑	SSCD↓	AES↑	CLIP↑
W/O mitigation	0.83	5.78	31.45	0.79	5.42	31.18	0.78	6.08	33.10
Deduplication	0.71	5.70	31.30	0.73	5.31	30.95	0.67	6.02	33.03
DP-LoRA	0.69	5.55	30.90	0.71	5.18	30.42	0.65	5.76	32.41
RTA	0.76	5.81	31.32	0.73	5.45	30.98	0.71	6.05	32.82
Prompt Optim.	0.72	5.63	31.05	0.74	5.28	30.70	0.69	5.88	32.58
CFG-Reverse	0.82	5.75	31.58	0.80	5.47	31.44	0.80	6.07	33.18
AMG	0.46	5.82	31.61	0.48	5.51	31.52	0.45	6.10	33.05
BAF-Soft	0.52	**5.97**	**33.52**	0.54	**6.01**	**34.23**	0.52	**6.35**	**34.58**
BAF-Hard	**0.49**	5.77	33.47	**0.52**	5.89	34.19	**0.50**	6.21	34.44

CFG-Reverse (Jain et al., 2025), and **AMG** (Chen et al., 2024b). These methods are post-training but require control over sampling; AMG additionally assumes access to training images and is included as a strong data-dependent reference.

5.1 Main Experiments

Table 1 and Table 2 report results across three datasets (Pokémon, CelebA-HQ, and LAION-Art) and three diffusion backbones. Across all settings, both BAF variants consistently achieve the lowest SSCD among *fairly comparable* methods, and even achieve better results than several methods that require additional information. We note that AMG achieves lower SSCD than the proposal because it suppresses memorization through similarity detection against the training set. However, this mechanism requires access to training images and relies on nearest-neighbor retrieval in the training embedding space during generation, introducing additional data dependency.

Importantly, BAF reduces memorization without sacrificing generation quality. Due to space constraints, we provide visualizations of the generated images with or without BAF in Appendix E. Across most settings, both AES and CLIP scores improve compared to the unfiltered LoRA. This suggests that weakly anchored channels may correspond to noisy parameter directions, and removing them can improve generation quality. BAF-Hard and BAF-Soft provide different trade-offs. BAF-Hard achieves stronger memorization suppression, while BAF-Soft preserves better generation quality. In practice, BAF-Soft provides a more balanced trade-off, and we use it in the following analyses.

5.2 Real-World Simulation on Civitai LoRAs

The previous experiments evaluate BAF under controlled memorization by artificially duplicating training images. To better reflect practical deployment, we further simulate a realistic post-hoc filtering scenario on publicly shared LoRAs. Specifically, we directly download LoRAs from Civitai and apply BAF without any access to the original training data or fine-tuning process, mirroring the setting faced by hosting platforms and downstream users. We evaluate BAF on 20 publicly available



Figure 3: Qualitative comparison on representative real-world LoRAs downloaded from Civitai. All methods use identical prompts and random seeds. BAF suppresses memorized identities while preserving the learned visual style.

LoRAs containing copyright-sensitive content, comparing it against RTA, CFG-Reverse, and Prompt Optim under identical prompts and random seeds.

Qualitative comparison. Due to space limitations, we first show two representative examples. Figure 3 presents results on two LoRAs downloaded from Civitai: a Goku toy-style LoRA on Stable Diffusion v1.5 and a clay-style LoRA on SDXL. In both cases, existing inference-time methods alter the outputs but largely preserve the memorized identity. In contrast, BAF produces visually distinct subjects while retaining the intended toy or clay style, suggesting better separation of reusable stylistic information from memorized identity. We further conduct a blind user study on all 20 community-released LoRAs with ten independent raters. BAF is consistently preferred for suppressing memorized identity while preserving the intended LoRA style. Full qualitative results and human evaluation details are provided in Appendix E.

5.3 Additional Analyses

We summarize three additional analyses here, with full quantitative results and visualizations provided in Appendix F. **(1) Effect of the soft-filtering strength.** Increasing α from 0.5 to 1 substantially improves memorization suppression, while larger values provide only marginal additional gains with little change in AES. We therefore use $\alpha = 1$ as the default. **(2) Effect of principal-subspace selection.** Varying τ_{energy} changes the pretrained principal subspace and consequently the anchoring threshold a_{null}. Across both CelebA-HQ and Pokémon, $\tau_{\text{energy}} = 0.85$ achieves the best memorization–utility trade-off. The corresponding anchoring-score distributions show a clearer separation between memorized and reusable adaptation, further supporting our anchoring hypothesis. **(3) Robustness to memorization strength.** We vary the canary duplication factor to induce different levels of memorization. Although higher duplication substantially increases memorization before filtering, BAF consistently reduces both targeted SSCD and Top-5% SSCD across all settings.

6 Conclusion

In this work, we proposed **BAF (Base-Anchored Filtering)**, a training-free, data-free method for mitigating memorization directly from released diffusion LoRA adapters. By leveraging the spectral geometry of the pretrained backbone, BAF separates memorization from generalizable adaptation through weight-space filtering. Extensive experiments on both controlled benchmarks and real-world LoRAs show that BAF effectively mitigates memorization while preserving generation quality.

Our results further suggest that parameter-space geometry provides useful structural signals for distinguishing memorization from generalizable adaptation. We hope this perspective inspires future research on parameter-space safety for foundation models and other PEFT methods.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, including grammar correction, improving readability, and condensing text to fit within reasonable space constraints while preserving the intended meaning. We also used generative AI tools to assist with extending the experimental implementation. The initial implementation was developed by the authors, and AI tools were subsequently used to help adapt and extend the codebase to support experiments across multiple models and configurations. In addition, we used generative AI tools as an auxiliary check on the correctness and consistency of our theoretical derivations and arguments. All AI-assisted code was manually reviewed, tested, and verified by the authors, and all theoretical claims and derivations were independently reviewed and validated by the authors. We did not use generative AI tools for other tasks requiring disclosure, such as developing the research methodology, formulating hypotheses, or interpreting experimental results. We take full responsibility for the final manuscript, implementation, theoretical results, experimental results, and all claims presented in this work.

ETHICS STATEMENT

This work studies memorization in diffusion LoRAs with the goal of mitigating unintended reproduction of copyrighted, private, or otherwise sensitive training content. Our controlled experiments use benchmark datasets and intentionally induced memorization solely for evaluating mitigation effectiveness. For the real-world evaluation, we use publicly available LoRAs released on Civitai and do not attempt to recover or redistribute their original training data. Qualitative examples are included only to demonstrate memorization behavior and the effectiveness of the proposed mitigation method.

Our human evaluation assesses generated images for memorized-identity suppression and preservation of the intended LoRA style. Participants were informed of the evaluation task and their responses were used only in aggregate. We recognize that identifying memorization in model parameters could potentially be misused to analyze or extract memorized information. However, our method is designed specifically for post-hoc suppression of such information and does not provide a procedure for recovering memorized training examples. We release our findings to support safer sharing and deployment of parameter-efficient adaptations. We take responsibility for the ethical conduct and presentation of the experiments in this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of our results. The complete BAF procedure is described in Appendix B. Experimental settings, including datasets, LoRA training configurations, diffusion backbones, evaluation metrics, and baseline configurations, are documented in Appendix D. We also include the code used for our experiments in the supplementary material.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7319–7328, 2021.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Ruchika Chavhan, Ondrej Bohdal, Yongshuo Zong, Da Li, and Timothy Hospedales. Memorized images in diffusion models share a subspace that can be located and deleted. *arXiv preprint arXiv:2406.18566*, 2024.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Exploring local memorization in diffusion models via bright ending attention. *arXiv preprint arXiv:2410.21665*, 2024a.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8425–8434, 2024b.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.

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*, pp. 2426–2436, 2023.

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*, pp. 5111–5120, 2024.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: low-rank adaptation of large language models. arxiv preprint. *arXiv preprint arXiv:2106.09685*, 2021.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12871–12879, 2025.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

LAION-AI. Laion-aesthetics. https://github.com/LAION-AI/aesthetic-predictor, 2022.

OpenAI. Dall·e 2 system card. https://openai.com/index/dall-e-2-pre-training-mitigations/, 2022. Accessed: 2026-02-23.

Justin N. M. Pinkney. Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2024.

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*, pp. 10684–10695, 2022.

Brendan Leigh Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse C Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. *arXiv preprint arXiv:2411.00113*, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4733–4743, 2024.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Yu-Lin Tsai, Yizhe Li, Chia-Mu Yu, Xuebin Ren, Po-Yu Chen, Zekai Chen, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4561–4571, 2025.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

A Details of the Directional Ablation Experiment

In the main paper, we design experiments to empirically validate the core hypothesis of the proposed method: that generalizable adaptation concentrates in spectral channels with high alignment to the pretrained model's principal subspace, while memorized content concentrates in weakly aligned channels. Due to space limitations, we are unable to describe all experimental settings in detail in the main paper. Therefore, we provide the detailed experimental settings here.

Setup: Two datasets are used in these experiments, Pokémon Pinkney (2022) and CelebA-HQ Karras et al. (2018). We used Stable Diffusion V1.5 as the pretrained backbone. LoRA updates are applied to the U-Net attention projection layers.

LoRA Training and Memorization simulation: To reliably induce and control memorization, two LoRA variants are trained under the same hyperparameter configuration.

High-mem LoRA (memorizing variant). A set of $N_c = 50$ canary images $\{(\mathbf{x}_k, y_k)\}_{k=1}^{N_c}$ is selected from the training data. Each canary image is duplicated $D = 50$ times within the training set, artificially amplifying memorization pressure on these specific samples. The resulting LoRA is expected to exhibit strong, measurable memorization of the canary images.

Low-mem LoRA (control variant). LoRAs are trained on the *same dataset without any canary image duplication*. This variant serves as a low-memorization reference baseline, representing standard fine-tuning conditions without deliberate memorization amplification.

We set the learning rate as 1e-4, optimizer as AdamW, training steps as 2000, LoRA rank as 8. All these training hyperparameters are held constant between the two variants.

The following procedure is applied *independently to each layer* of the High-mem LoRA.

Step 1: SVD of the LoRA update. For a layer with pretrained weight $\mathbf{W}^* \in \mathbb{R}^{m\times n}$ and LoRA update $\Delta W = BA \in \mathbb{R}^{m\times n}$ of rank r, the SVD decomposes ΔW into spectral channels:

$$\Delta W = \sum_{i=1}^{r} \sigma_i \, \mathbf{u}_i \mathbf{v}_i^\top, \tag{6}$$

where $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are singular values and $\mathbf{u}_i \in \mathbb{R}^m$, $\mathbf{v}_i \in \mathbb{R}^n$ are the corresponding left and right singular vectors.

Step 2: Anchoring score computation. Let $\mathbf{P}_L = \mathbf{U}_{*,K}\mathbf{U}_{*,K}^\top$ and $\mathbf{P}_R = \mathbf{V}_{*,K}\mathbf{V}_{*,K}^\top$ denote orthogonal projections onto the top-K left and right principal subspaces of $\mathbf{W}^*$, respectively. The **anchoring score** for the i-th spectral channel is defined as:

$$a_i = \|\mathbf{P}_L \mathbf{u}_i\|^2 \cdot \|\mathbf{P}_R \mathbf{v}_i\|^2 = \cos^2\theta_i^L \cdot \cos^2\theta_i^R \ \in [0,1], \tag{7}$$

where θ_i^L and θ_i^R are the angles between $\mathbf{u}_i$, $\mathbf{v}_i$ and the respective pretrained subspaces. The subspace dimension K is selected via the energy-adaptive rule with threshold $\tau_{\text{energy}} = 0.85$.

Step 3: Random-projection baseline. To establish a reference for alignment, the layer-wise null baseline is:

$$a_{\text{null}} = \frac{K^2}{mn}. \tag{8}$$

This baseline equals the expected anchoring score of a channel whose singular vectors bear no structured relationship to the pretrained model.

Step 4: Binary channel splitting. Based on the anchoring score, we construct two complementary ablated LoRA variants.

High-align-only. All spectral channels satisfying $a_i \geq a_{\text{null}}$ are retained, while all low-alignment channels ($a_i < a_{\text{null}}$) are removed. This variant isolates the highly anchored directions that are expected to capture generalizable adaptation.

Low-align-only. Conversely, all spectral channels satisfying $a_i < a_{\text{null}}$ are retained, while the high-alignment channels are removed. This variant isolates the weakly anchored directions that are hypothesized to contain memorized content.

Formally, each ablated variant is obtained by re-weighting the original spectral channels:

$$\begin{aligned} \Delta W_{\text{ablated}} &= \sum_{i=1}^{r} \sigma_i \, g(a_i) \, \mathbf{u}_i \mathbf{v}_i^\top, \\ g(a_i) &= \begin{cases} \mathbf{1}[a_i \geq a_{\text{null}}], & \text{(High-align-only)}, \\ \mathbf{1}[a_i < a_{\text{null}}], & \text{(Low-align-only)}. \end{cases} \end{aligned} \tag{9}$$

Algorithm 1 Base-Anchored Filtering (BAF)

Require: Pretrained weights $\{W_*^{(\ell)}\}_{\ell=1}^{L}$; LoRA factors $\{B^{(\ell)}, A^{(\ell)}\}_{\ell=1}^{L}$; energy threshold $\tau_{\text{energy}} \in (0,1)$; gating mode $\text{mode} \in \{\text{HARD}, \text{SOFT}\}$; softness exponent $\alpha > 0$ for SOFT mode

Ensure: Filtered LoRA updates $\{\Delta W_{\text{BAF}}^{(\ell)}\}_{\ell=1}^{L}$

for $\ell = 1$ **to** L **do**
  {Stage 1: Spectral decomposition of pretrained weight}
  $[\mathbf{U}_*, \mathbf{\Sigma}_*, \mathbf{V}_*] \leftarrow \text{SVD}(W_*^{(\ell)})$
  {Stage 2: Energy-adaptive subspace selection}
  $K \leftarrow \min\left\{k : \frac{\sum_{j=1}^{k} s_j^2}{\sum_{j=1}^{\min(m,n)} s_j^2} \geq \tau_{\text{energy}}\right\}$
  $\mathbf{U}_{*,K} \leftarrow \mathbf{U}_*[:, 1{:}K]$
  $\mathbf{V}_{*,K} \leftarrow \mathbf{V}_*[:, 1{:}K]$
  $\mathbf{P}_L \leftarrow \mathbf{U}_{*,K}\mathbf{U}_{*,K}^{\top}$
  $\mathbf{P}_R \leftarrow \mathbf{V}_{*,K}\mathbf{V}_{*,K}^{\top}$
  $a_{\text{null}} \leftarrow K^2/(mn)$
  {Stage 3: Spectral decomposition of LoRA update}
  $\Delta W^{(\ell)} \leftarrow B^{(\ell)}A^{(\ell)}$
  $[\mathbf{U}, \boldsymbol{\sigma}, \mathbf{V}] \leftarrow \text{SVD}(\Delta W^{(\ell)})$
  {Stage 4: Anchoring-score computation}
  for $i = 1$ **to** r **do**
    $a_i \leftarrow \|\mathbf{P}_L\mathbf{u}_i\|_2^2 \, \|\mathbf{P}_R\mathbf{v}_i\|_2^2$
  end for
  {Stage 5: Spectral gating}
  if $\text{mode} = \text{HARD}$ **then**
    for $i = 1$ **to** r **do**
      $g(a_i) \leftarrow \mathbf{1}[a_i \geq a_{\text{null}}]$
    end for
  else
    for $i = 1$ **to** r **do**
      $g(a_i) \leftarrow a_i^{\alpha}$
    end for
  end if
  {Stage 6: Filtered-update reconstruction}
  $\Delta W_{\text{BAF}}^{(\ell)} \leftarrow \sum_{i=1}^{r} \sigma_i g(a_i) \mathbf{u}_i \mathbf{v}_i^{\top}$
end for
return $\{\Delta W_{\text{BAF}}^{(\ell)}\}_{\ell=1}^{L}$

Evaluation Metrics Two complementary metrics are reported to separately assess memorization and generation quality.

- **Memorization:** Cosine similarity in **CLIP embedding space** between generated images and the corresponding canary training images. A higher similarity score indicates stronger memorization of the training data.
- **Generation quality: Aesthetic Score (AES)** computed by the LAION-Aesthetics predictor LAION-AI (2022), measuring the perceptual visual quality of generated outputs.

B FULL ALGORITHM

We provide the full algorithm box in Algorithm 1.

C THEORETICAL ANALYSIS

The main paper presents a simplified version of the theoretical result to highlight the key intuition behind BAF and improve readability. This appendix provides the complete theoretical formulation together with detailed derivations and proofs.

Specifically, we first introduce a counterfactual decomposition of a released LoRA update into clean adaptation and memorization components. Based on this decomposition, we derive an exact recovery-error identity that holds without additional assumptions. We then introduce measurable energy-leakage quantities that quantify the separation assumptions used in the main paper, and establish deterministic recovery guarantees for both BAF-Hard and BAF-Soft. Finally, we empirically validate the quantities appearing in the recovery bounds and examine whether the derived upper bounds are consistent with the observed recovery error in practice.

C.1 SETUP AND COUNTERFACTUAL DECOMPOSITION

We analyze a single LoRA-adapted weight matrix and suppress the layer index throughout this section. Let

$$\boldsymbol{W}_* \in \mathbb{R}^{m\times n}$$

denote the corresponding pretrained weight matrix. Let $\boldsymbol{P}_L$ and $\boldsymbol{P}_R$ denote the orthogonal projectors onto the energy-adaptive left and right principal subspaces of $\boldsymbol{W}_*$, respectively.

The released LoRA update admits the singular-value decomposition

$$\Delta\boldsymbol{W} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \tag{10}$$

where the rank-one matrices $\{\boldsymbol{u}_i\boldsymbol{v}_i^\top\}_{i=1}^r$ form an orthonormal basis of the rank-r singular-channel subspace

$$\mathcal{S} := \operatorname{span}\left\{\boldsymbol{u}_i\boldsymbol{v}_i^\top\right\}_{i=1}^r.$$

Since BAF acts independently on each singular channel, all subsequent analysis is performed in this subspace.

The anchoring score of the ith channel is

$$a_i = \|\boldsymbol{P}_L\boldsymbol{u}_i\|_2^2\,\|\boldsymbol{P}_R\boldsymbol{v}_i\|_2^2. \tag{11}$$

For an arbitrary gate $\boldsymbol{g} = (g_1,\dots,g_r) \in [0,1]^r$, the filtered update is

$$\Delta\boldsymbol{W}(\boldsymbol{g}) = \sum_{i=1}^{r} g_i\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top. \tag{12}$$

BAF-Hard uses

$$g_i = \mathbf{1}[a_i \ge a_{\text{null}}],$$

whereas BAF-Soft uses

$$g_i = a_i^\alpha.$$

Counterfactual decomposition. Let $\Delta\boldsymbol{W}_{\text{clean}}$ denote the LoRA update obtained under the counterfactual training set without duplicated examples. We define the duplication-induced residual as

$$\Delta\boldsymbol{W}_{\text{mem}} := \Delta\boldsymbol{W} - \Delta\boldsymbol{W}_{\text{clean}}, \tag{13}$$

so that

$$\Delta\boldsymbol{W} = \Delta\boldsymbol{W}_{\text{clean}} + \Delta\boldsymbol{W}_{\text{mem}}. \tag{14}$$

This decomposition is purely definitional. It does not assume that LoRA training is linear in the training data, nor that clean adaptation and memorization are independently learned. Instead, $\Delta\boldsymbol{W}_{\text{mem}}$ absorbs every parameter-space change caused by duplication, including nonlinear optimization effects and interactions with clean adaptation.

Projection onto the singular-channel subspace. Because BAF only modifies the singular channels of $\Delta \boldsymbol{W}$, only the components lying in $\mathcal{S}$ are directly controllable.

Let $\Pi_{\mathcal{S}}(\cdot)$ denote the orthogonal projection onto $\mathcal{S}$. Projecting the clean and duplication-induced components gives

$$\Pi_{\mathcal{S}}(\Delta \boldsymbol{W}_{\text{clean}}) = \sum_{i=1}^{r} c_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \qquad \Pi_{\mathcal{S}}(\Delta \boldsymbol{W}_{\text{mem}}) = \sum_{i=1}^{r} m_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \tag{15}$$

where

$$c_i = \left\langle \Delta \boldsymbol{W}_{\text{clean}}, \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\rangle_{\text{F}}, \qquad m_i = \left\langle \Delta \boldsymbol{W}_{\text{mem}}, \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\rangle_{\text{F}}. \tag{16}$$

Since $\Delta \boldsymbol{W} = \Delta \boldsymbol{W}_{\text{clean}} + \Delta \boldsymbol{W}_{\text{mem}}$, linearity of the Frobenius inner product gives

$$c_i + m_i = \sigma_i. \tag{17}$$

The projection does not capture components orthogonal to $\mathcal{S}$. Accordingly, we write

$$\Delta \boldsymbol{W}_{\text{mem}} = \Pi_{\mathcal{S}}(\Delta \boldsymbol{W}_{\text{mem}}) + \boldsymbol{R}, \tag{18}$$

where

$$\boldsymbol{R} \perp \mathcal{S}.$$

Since the released update $\Delta \boldsymbol{W}$ lies entirely in $\mathcal{S}$, the orthogonal residuals of the two counterfactual components must cancel, yielding

$$\Delta \boldsymbol{W}_{\text{clean}} = \Pi_{\mathcal{S}}(\Delta \boldsymbol{W}_{\text{clean}}) - \boldsymbol{R}. \tag{19}$$

Finally, we characterize the fraction of duplication-induced energy that cannot be represented in the singular-channel subspace by

$$\kappa = \frac{\|\boldsymbol{R}\|_{\text{F}}}{\|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}}, \tag{20}$$

with $\kappa = 0$ when $\|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}} = 0$. By construction, $0 \le \kappa \le 1$.

C.2 Geometric Decomposition of the Recovery Error

The projection view introduced above naturally separates the recovery error into two components: one inside the singular-channel subspace $\mathcal{S}$, which can be modified by channel-wise filtering, and one orthogonal to $\mathcal{S}$, which is fundamentally beyond the reach of BAF. The following proposition formalizes this decomposition.

Proposition 1 (Geometric recovery-error decomposition)**.** *For any channel-wise gate $\boldsymbol{g} \in [0,1]^r$,*

$$\begin{aligned} \|\Delta \boldsymbol{W}(\boldsymbol{g}) - \Delta \boldsymbol{W}_{\text{clean}}\|_{\text{F}}^2 = &\sum_{i=1}^{r} \left(g_i m_i - (1-g_i) c_i\right)^2 \\ &+ \kappa^2 \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}^2. \end{aligned} \tag{21}$$

Proof. Using Eq. equation 18 and Eq. equation 19,

$$\begin{aligned}\Delta \boldsymbol{W}(\boldsymbol{g}) - \Delta \boldsymbol{W}_{\text{clean}} &= \sum_{i=1}^{r} g_i(c_i + m_i)\, \boldsymbol{u}_i \boldsymbol{v}_i^\top - \Delta \boldsymbol{W}_{\text{clean}} \\ &= \sum_{i=1}^{r} \left(g_i(c_i + m_i) - c_i\right) \boldsymbol{u}_i \boldsymbol{v}_i^\top + \boldsymbol{R} \\ &= \sum_{i=1}^{r} \left(g_i m_i - (1 - g_i)c_i\right) \boldsymbol{u}_i \boldsymbol{v}_i^\top + \boldsymbol{R}. \end{aligned} \tag{22}$$

The first term lies entirely in the singular-channel subspace $\mathcal{S}$, whereas $\boldsymbol{R} \perp \mathcal{S}$ by construction. Applying the Pythagorean theorem therefore gives

$$\|\Delta \boldsymbol{W}(\boldsymbol{g}) - \Delta \boldsymbol{W}_{\text{clean}}\|_{\text{F}}^2 = \sum_{i=1}^{r} \left(g_i m_i - (1 - g_i)c_i\right)^2 + \|\boldsymbol{R}\|_{\text{F}}^2.$$

Finally, $\|\boldsymbol{R}\|_{\text{F}} = \kappa \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}$ from Eq. equation 20, which proves Eq. equation 21.

□

Proposition 1 reveals that the recovery error has three distinct sources.

First, the term $g_i m_i$ measures memorization retained after filtering. A stronger gate suppresses this contribution.

Second, the term $(1 - g_i)c_i$ measures clean adaptation discarded by filtering. Reducing memorization therefore inevitably competes with preserving useful adaptation.

Finally, the residual term κ corresponds to the duplication-induced component outside the singular-channel subspace. Unlike the first two terms, this component is independent of the gating function. Since BAF operates exclusively within $\mathcal{S}$, it cannot directly modify this residual.

The decomposition therefore separates the controllable and uncontrollable parts of recovery: the gate determines the trade-off between retained memorization and preserved clean adaptation inside $\mathcal{S}$, whereas κ represents an irreducible recovery floor determined by the geometry of the released update.

Special case: BAF-Hard. For BAF-Hard, each singular channel is either completely preserved or completely removed. Consequently, retained memorization and discarded clean adaptation never coexist within the same channel. Equation equation 21 therefore simplifies to

$$\begin{aligned}\|\Delta \boldsymbol{W}_{\text{hard}}(\tau) - \Delta \boldsymbol{W}_{\text{clean}}\|_{\text{F}}^2 = &\sum_{i:a_i \geq \tau} m_i^2 + \sum_{i:a_i < \tau} c_i^2 \\ &+ \kappa^2 \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}^2, \end{aligned} \tag{23}$$

where τ denotes the hard gating threshold.

Equation equation 23 shows that, under hard gating, the recovery error is exactly the sum of

(i) memorization energy retained in the selected channels,

(ii) clean adaptation removed from the discarded channels,

and

(iii) the irreducible residual outside the singular-channel subspace.

C.3 Approximate Recovery Under Energy Separation

Theorem 1 in the main paper states that BAF approximately recovers the counterfactual clean LoRA whenever memorized content and generalizable adaptation are well separated in the anchoring-score space. This section provides a formal proof of that result.

The key observation is that Proposition 1 already gives an exact decomposition of the recovery error without requiring any assumption. The remaining question is therefore how large each term in that decomposition can become.

Instead of treating Assumptions 1 and 2 as qualitative statements, we introduce two measurable quantities that quantify the extent to which these assumptions are satisfied by a given LoRA checkpoint.

Definition 3 (Memorization leakage)**.** For a threshold $s \in (0, 1)$, define

$$\rho_{\mathrm{mem}}(s)^2 := \frac{\sum_{i:a_i \geq s} m_i^2}{\|\Delta \boldsymbol{W}_{\mathrm{mem}}\|_{\mathrm{F}}^2}. \tag{24}$$

The quantity $\rho_{\mathrm{mem}}(s)$ measures the fraction of duplication-induced energy remaining inside strongly anchored channels. A small value therefore indicates that memorization is concentrated in weakly anchored channels, consistent with Assumption 2.

Definition 4 (Clean-energy leakage)**.** For $\epsilon \in [0, 1)$, define

$$\rho_{\mathrm{clean}}(\epsilon)^2 := \frac{\sum_{i:a_i < 1-\epsilon} c_i^2}{\|\Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}}^2}. \tag{25}$$

The quantity $\rho_{\mathrm{clean}}(\epsilon)$ measures the fraction of clean adaptation energy outside the strongly anchored region. A small value therefore indicates that most useful adaptation is preserved after filtering, consistent with Assumption 1.

Unlike qualitative assumptions, both leakage coefficients are directly measurable from paired clean and memorization-inducing LoRA checkpoints. Consequently, once ρ_{clean}, ρ_{mem}, and κ are measured, the recovery guarantee becomes fully deterministic.

Theorem 2 (Approximate recovery)**.** *Let*

$$\varepsilon(\boldsymbol{g}) := \|\Delta \boldsymbol{W}(\boldsymbol{g}) - \Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}}.$$

Then the following recovery guarantees hold.

BAF-Hard. *Suppose the hard gate uses threshold $\tau \leq 1 - \epsilon$. Then*

$$\begin{aligned} \varepsilon_{\mathrm{hard}}(\tau)^2 \leq \left(\rho_{\mathrm{mem}}(\tau)^2 + \kappa^2\right) \|\Delta \boldsymbol{W}_{\mathrm{mem}}\|_{\mathrm{F}}^2 \\ + \rho_{\mathrm{clean}}(\epsilon)^2 \|\Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}}^2. \end{aligned} \tag{26}$$

BAF-Soft. *For any $\alpha > 0$,*

$$\begin{aligned} \varepsilon_{\mathrm{soft}}(\alpha) \leq & \eta_{\mathrm{clean}}(\alpha, \epsilon) \|\Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}} \\ & + \left(\eta_{\mathrm{mem}}(\alpha, s) + \kappa\right) \|\Delta \boldsymbol{W}_{\mathrm{mem}}\|_{\mathrm{F}}, \end{aligned} \tag{27}$$

where

$$\eta_{\mathrm{clean}}(\alpha, \epsilon) = \sqrt{\left(1 - (1-\epsilon)^{\alpha}\right)^2 + \rho_{\mathrm{clean}}(\epsilon)^2}, \tag{28}$$

and

$$\eta_{\mathrm{mem}}(\alpha, s) = \sqrt{\rho_{\mathrm{mem}}(s)^2 + s^{2\alpha}}. \tag{29}$$

Theorem 2 is the formal version of the approximate recovery result presented in the main paper.

Remark 4 (Correspondence to Theorem 1 in the main paper). Theorem 1 in the main paper is a simplified presentation of Theorem 2 intended to emphasize the main intuition. For readability, several notation simplifications are adopted there, whereas this appendix uses the fully explicit quantities required for the complete proof.

First, the symbol κ in the main paper denotes the *absolute* irreducible recovery error,

$$\kappa_{\text{main}} = \kappa \, \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}},$$

whereas throughout this appendix $\kappa \in [0, 1]$ denotes the normalized geometric misfit defined in Eq. equation 20.

Second, the leakage coefficients appearing in the main paper are shorthand notations

$$\rho_c := \rho_{\text{clean}}(\epsilon), \qquad \rho_m := \rho_{\text{mem}}(s),$$

with the thresholds (ϵ, s) omitted for simplicity.

Finally, the constants C_1 and C_2 in the main paper absorb the gating-dependent factors appearing in the detailed bounds. For BAF-Hard, $C_1 = C_2 = 1$, and Eq. equation 26 reduces to the simplified statement in the main paper by applying $\sqrt{x^2 + y^2 + z^2} \le x + y + z$. For BAF-Soft,

$$C_1 := \frac{\eta_{\text{clean}}(\alpha, \epsilon)}{\rho_{\text{clean}}(\epsilon)}, \qquad C_2 := \frac{\eta_{\text{mem}}(\alpha, s)}{\rho_{\text{mem}}(s)},$$

which explicitly depend on the filtering parameter α and the thresholds (ϵ, s).

Throughout this appendix, all derivations and proofs use the explicit formulation given in Theorem 2. The simplified theorem in the main paper is obtained directly by substituting the above identifications.

The recovery error consists of three components. The first is the leakage of useful adaptation into weakly anchored channels, quantified by ρ_{clean}. The second is the leakage of memorization into strongly anchored channels, quantified by ρ_{mem}. The third is the intrinsic geometric mismatch κ, which measures duplication-induced energy lying outside the singular-channel span of the released update.

Therefore, BAF approximately recovers the counterfactual clean LoRA whenever the two leakage coefficients remain small. The remaining recovery error is bounded by the irreducible overlap term κ, explaining why the theorem establishes approximate rather than exact recovery.

Proof. The result follows by combining the exact decomposition in Proposition 1 with the definitions of the energy-leakage coefficients.

BAF-Hard. For hard gating,

$$g_i = \mathbf{1}[a_i \ge \tau],$$

and Proposition 1 reduces to

$$\begin{aligned} \varepsilon_{\text{hard}}(\tau)^2 = \sum_{i:a_i \ge \tau} m_i^2 + \sum_{i:a_i < \tau} c_i^2 \\ + \kappa^2 \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}^2. \end{aligned} \tag{30}$$

By Definition 3,

$$\sum_{i:a_i \ge \tau} m_i^2 = \rho_{\text{mem}}(\tau)^2 \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}^2.$$

Moreover,

$$\tau \le 1 - \epsilon$$

implies

$$\{a_i < \tau\} \subseteq \{a_i < 1 - \epsilon\},$$

and therefore

$$\sum_{i:a_i<\tau} c_i^2 \le \rho_{\text{clean}}(\epsilon)^2 \|\Delta \boldsymbol{W}_{\text{clean}}\|_{\text{F}}^2.$$

Substituting these bounds into Eq. equation 30 proves Eq. equation 26.

BAF-Soft. For soft gating,

$$g_i = a_i^\alpha.$$

Applying Proposition 1 and the triangle inequality gives

$$\begin{aligned}\varepsilon_{\text{soft}}(\alpha) \le &\left(\sum_{i=1}^{r} a_i^{2\alpha} m_i^2\right)^{1/2} \\ &+ \left(\sum_{i=1}^{r} (1 - a_i^\alpha)^2 c_i^2\right)^{1/2} \\ &+ \kappa \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}.\end{aligned} \tag{31}$$

We first bound the memorization term.

Splitting the channels at threshold s,

$$\begin{aligned}\sum_i a_i^{2\alpha} m_i^2 &= \sum_{a_i \ge s} a_i^{2\alpha} m_i^2 + \sum_{a_i < s} a_i^{2\alpha} m_i^2 \\ &\le \sum_{a_i \ge s} m_i^2 + s^{2\alpha} \sum_{a_i < s} m_i^2 \\ &\le \left(\rho_{\text{mem}}(s)^2 + s^{2\alpha}\right) \|\Delta \boldsymbol{W}_{\text{mem}}\|_{\text{F}}^2.\end{aligned} \tag{32}$$

Next, consider the clean-adaptation term.

Splitting the channels at $1 - \epsilon$,

$$\begin{aligned}&\sum_i (1 - a_i^\alpha)^2 c_i^2 \\ &= \sum_{a_i \ge 1-\epsilon} (1 - a_i^\alpha)^2 c_i^2 + \sum_{a_i < 1-\epsilon} (1 - a_i^\alpha)^2 c_i^2 \\ &\le \left(1 - (1-\epsilon)^\alpha\right)^2 \sum_{a_i \ge 1-\epsilon} c_i^2 + \sum_{a_i < 1-\epsilon} c_i^2 \\ &\le \left[\left(1 - (1-\epsilon)^\alpha\right)^2 + \rho_{\text{clean}}(\epsilon)^2\right] \|\Delta \boldsymbol{W}_{\text{clean}}\|_{\text{F}}^2.\end{aligned} \tag{33}$$

Substituting Eq. equation 32 and Eq. equation 33 into Eq. equation 31 yields Eq. equation 27.

□

Remark 5 (Trade-off induced by the filtering strength). The recovery bound explicitly characterizes the role of the soft-gating parameter.

Increasing α reduces $s^{2\alpha}$, thereby suppressing memorization retained in weakly anchored channels.

However, larger α also increases

$$1 - (1 - \epsilon)^{\alpha},$$

meaning that even strongly anchored clean channels are attenuated more aggressively.

Therefore, stronger filtering improves memorization suppression at the cost of removing more useful adaptation, making the trade-off explicit in the recovery bound.

Remark 6 (Irreducible recovery floor). The quantity

$$\kappa \|\Delta \boldsymbol{W}_{\mathrm{mem}}\|_{\mathrm{F}}$$

is independent of the gating function.

It represents duplication-induced energy outside the singular-channel subspace $\mathcal{S}$, which cannot be modified by any channel-wise filtering strategy.

Consequently, even an ideal gating function cannot achieve exact recovery unless $\kappa = 0$. Theorem 2 therefore establishes approximate rather than exact recovery, with κ quantifying the unavoidable recovery floor.

C.4 From Weight Recovery to Behavioral Recovery

The recovery guarantees in Theorem 2 are stated entirely in parameter space. However, the quantities evaluated throughout the paper, including memorization and generation quality, are behavioral metrics rather than weight-space distances. This subsection establishes a bridge between the two.

Throughout this subsection, let

$$M(\Delta \boldsymbol{W})$$

denote the model obtained by applying the LoRA update $\Delta \boldsymbol{W}$ to the pretrained backbone. Define

$$M_{\mathrm{ori}} := M(\Delta \boldsymbol{W}), \qquad M_{\mathrm{clean}} := M(\Delta \boldsymbol{W}_{\mathrm{clean}}),$$

$$M_{\mathrm{BAF}} := M(\Delta \boldsymbol{W}_{\mathrm{BAF}}).$$

The following assumption states that, within the local neighborhood visited by BAF, the evaluation metrics vary smoothly with the LoRA update.

Assumption 1 (Local metric regularity)**.** There exist nonnegative constants L_m and L_u such that, for any two LoRA updates $\Delta \boldsymbol{W}_1$ and $\Delta \boldsymbol{W}_2$ inside the local neighborhood of interest,

$$|m(M(\Delta \boldsymbol{W}_1)) - m(M(\Delta \boldsymbol{W}_2))| \le L_m \|\Delta \boldsymbol{W}_1 - \Delta \boldsymbol{W}_2\|_{\mathrm{F}}, \tag{34}$$

$$|u(M(\Delta \boldsymbol{W}_1)) - u(M(\Delta \boldsymbol{W}_2))| \le L_u \|\Delta \boldsymbol{W}_1 - \Delta \boldsymbol{W}_2\|_{\mathrm{F}}, \tag{35}$$

where $m(\cdot)$ denotes a memorization metric (e.g., Targeted SSCD), and $u(\cdot)$ denotes a utility metric (e.g., CLIP score or aesthetic score).

The assumption does not require global smoothness. It only concerns the local parameter region traversed by BAF, which modifies the released LoRA update through relatively small channel-wise perturbations.

Corollary 1 (Behavioral recovery)**.** *Suppose*

$$\|\Delta \boldsymbol{W}_{\mathrm{BAF}} - \Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}} \leq \varepsilon,$$

where ε is any upper bound provided by Theorem 2.

Under Assumption 1,

$$|m(M_{\mathrm{BAF}}) - m(M_{\mathrm{clean}})| \leq L_m \varepsilon, \tag{36}$$
$$|u(M_{\mathrm{BAF}}) - u(M_{\mathrm{clean}})| \leq L_u \varepsilon. \tag{37}$$

Furthermore, if

$$m(M_{\mathrm{clean}}) + L_m \varepsilon < m(M_{\mathrm{ori}}), \tag{38}$$

then

$$m(M_{\mathrm{BAF}}) < m(M_{\mathrm{ori}}), \tag{39}$$

meaning that BAF provably reduces memorization relative to the original released LoRA.

Proof. Assumption 1 directly gives

$$|m(M_{\mathrm{BAF}}) - m(M_{\mathrm{clean}})| \leq L_m \|\Delta \boldsymbol{W}_{\mathrm{BAF}} - \Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}},$$

and similarly for the utility metric. Substituting

$$\|\Delta \boldsymbol{W}_{\mathrm{BAF}} - \Delta \boldsymbol{W}_{\mathrm{clean}}\|_{\mathrm{F}} \leq \varepsilon$$

yields Eqs. equation 36 and equation 37.

Finally,

Eq. equation 38 together with Eq. equation 36 implies

$$m(M_{\mathrm{BAF}}) \leq m(M_{\mathrm{clean}}) + L_m \varepsilon < m(M_{\mathrm{ori}}),$$

which proves Eq. equation 39.

□

Corollary 1 establishes a direct bridge between parameter-space recovery and behavioral recovery. Combined with Theorem 2, it shows that the effectiveness of BAF is governed by two factors.

First, small leakage coefficients ρ_{mem} and ρ_{clean} lead to a small parameter-space recovery error. Second, under local metric regularity, small parameter perturbations induce proportionally small changes in memorization and utility metrics. Consequently, recovering the clean counterfactual in parameter space also approximately recovers its memorization and utility behaviors.

The behavioral guarantee is intentionally local rather than global. Metrics such as Targeted SSCD, CLIP score, and aesthetic score are computed from highly nonlinear diffusion models and generally do not admit meaningful global Lipschitz constants. The local regularity assumption therefore serves as a standard bridge between the geometric recovery analysis and the empirical behaviors evaluated in the experiments.

Table 3: Empirical validation of the recovery-theoretic quantities on SD 1.5 LoRA adapters trained on the Pokémon memorization benchmark. The theoretical optimum corresponds to perfect recoverability, whereas the upper bound is determined by the definitions of the corresponding quantities.

Quantity	Measured	Theoretical Optimum	Upper Bound
ρ_{mem}	0.29	0	1
ρ_{clean}	0.22	0	1
κ	0.45	0	1

C.5 Empirical Validation of the Recovery Theory

The preceding analysis establishes that the recovery performance of BAF is determined by three measurable quantities: the memorization leakage coefficient ρ_{mem}, the clean-adaptation leakage coefficient ρ_{clean}, and the channel-span misfit κ. Theorem 2 predicts that successful recovery requires both leakage coefficients to remain close to their theoretical optimum of zero, while Proposition 1 identifies κ as the irreducible recovery floor.

This subsection examines whether these theoretical conditions are satisfied in practical diffusion LoRA adapters.

Experimental setting. We perform the validation using LoRA adapters trained on the Pokémon dataset with the SD 1.5 backbone. Following the memorization benchmark used throughout the paper, memorization is induced by repeatedly duplicating a subset of Pokémon training images while keeping all remaining training configurations unchanged. The corresponding clean LoRA is trained using the identical pipeline without duplication, providing the counterfactual reference $\Delta \boldsymbol{W}_{\mathrm{clean}}$. The duplication-induced component is therefore obtained as

$$\Delta \boldsymbol{W}_{\mathrm{mem}} = \Delta \boldsymbol{W} - \Delta \boldsymbol{W}_{\mathrm{clean}}.$$

For every adapted module, we compute the singular-value decomposition of the released LoRA update, evaluate the anchoring score of every singular channel, and estimate the theoretical quantities ρ_{mem}, ρ_{clean}, and κ according to Definitions 3, 4, and Eq. equation 20. The reported values are averaged across all adapted modules.

Table 3 compares the measured theoretical quantities with their corresponding theoretical limits.

The memorization leakage coefficient ρ_{mem} measures the fraction of duplication-induced energy retained inside the preserved spectral region. Its theoretical optimum is zero, indicating that all memorization lies in removable channels, whereas the upper bound of one corresponds to the least favorable recovery regime. The measured value, $\rho_{\mathrm{mem}} = 0.29$, is substantially closer to the theoretical optimum than to the upper bound, implying that approximately 71% of the duplication-induced energy lies outside the preserved channels and is therefore removable by channel-wise filtering.

Similarly, ρ_{clean} quantifies the fraction of clean adaptation discarded by filtering. Its measured value, $\rho_{\mathrm{clean}} = 0.22$, indicates that approximately 78% of the clean adaptation remains inside the preserved spectral region. This energy separation between memorization and clean adaptation is precisely the condition required by Theorem 2 for successful recovery.

Finally, the channel-span misfit κ measures the fraction of duplication-induced energy outside the singular-channel span of the released update. Unlike the leakage coefficients, κ does not characterize energy separation but the geometric limitation of channel-wise filtering. The measured value, $\kappa = 0.45$, is moderate rather than negligible, indicating that part of the memorization component cannot be represented within the singular-channel basis. This observation agrees with Proposition 1, which predicts an unavoidable recovery floor and therefore explains why BAF achieves approximate rather than exact recovery.

Overall, all measured quantities are substantially closer to their theoretical optima than to their corresponding upper bounds. These observations confirm that practical SD 1.5 LoRA adapters satisfy the geometric conditions assumed by the recovery theory. Combined with the behavioral results reported in the main paper, they provide empirical evidence for the theoretical prediction

that recovering the clean counterfactual in parameter space leads to simultaneous memorization suppression and utility preservation.

D Experimental Setup and Additional Quantitative Results

We describe the training hyperparameters used in the main experiments. Unless otherwise specified, the same configuration is used across datasets and base diffusion models.

Common settings. For all experiments we use a training batch size of 4 and set the maximum gradient norm to 1 for gradient clipping. The learning rate scheduler is fixed to a constant schedule. The training resolution follows the default resolution of the underlying diffusion backbone: 512×512 for Stable Diffusion 1.5, 768×768 for Stable Diffusion 2.1, 1024×1024 for SDXL, and 1024×1024 for FLUX.

Different diffusion backbones use their standard sampling schedulers. Specifically, Stable Diffusion 1.5 uses the PNDM scheduler, Stable Diffusion 2.1 uses the DPM-Solver multistep scheduler, SDXL uses the Euler discrete scheduler, and FLUX uses the FlowMatch Euler discrete scheduler.

CelebA-HQ training. For the CelebA-HQ dataset we train LoRA modules for 20 epochs with a learning rate of 10^{-4}. For Stable Diffusion 1.5 and Stable Diffusion 2.1 we use a LoRA rank of 8, while SDXL and FLUX both use a rank of 16.

Pokémon training. For the Pokémon dataset we use slightly different configurations. For Stable Diffusion 1.5 we train LoRA with rank 16 for 10 epochs using a learning rate of 10^{-4}. For Stable Diffusion 2.1 we also use rank 16 but train for 20 epochs with the same learning rate of 10^{-4}. For SDXL we use a rank of 8 and train for 15 epochs with a learning rate of 5×10^{-5}. For FLUX we use a rank of 16 and train for 15 epochs with a learning rate of 5×10^{-5}.

Laion-art training. For the Laion-art dataset we also use slightly different configurations. For Stable Diffusion 1.5 we train LoRA with rank 16 for 20 epochs using a learning rate of 10^{-4}. For Stable Diffusion 2.1 we also use rank 16 and train for 20 epochs with the same learning rate of 10^{-4}. For SDXL we use a rank of 16 and train for 30 epochs with a learning rate of 5×10^{-4}. For FLUX we also use a rank of 16 and train for 30 epochs with a learning rate of 5×10^{-5}.

Baseline For the baselines DP-LoRA, RTA, Prompt Optimization, CFG-Reverse, and AMG, we adopt the official implementations provided in the authors' public GitHub repositories. All hyperparameters are kept consistent with the configurations recommended by the original papers whenever possible. This ensures that each baseline is evaluated under settings representative of its intended usage. We also note that several baselines considered in this work were not originally proposed in the context of LoRA-based customization. In particular, CFG-Reverse, RTA, and Prompt Optimization were introduced for general diffusion models rather than LoRA-adapted models. Among them, RTA and Prompt Optimization operate at the prompt or generation level and do not depend on the internal parameterization of the model. Therefore, their effectiveness is largely independent of whether the model is obtained via LoRA adaptation or standard fine-tuning. In contrast, CFG-Reverse modifies the classifier-free guidance mechanism during inference and implicitly relies on the behavior of the underlying model parameters. As a result, its effectiveness may be affected when the model is adapted through LoRA updates. We nevertheless include CFG-Reverse for completeness, as it represents a commonly used inference-time strategy for reducing memorization-related artifacts.

To simulate a realistic training scenario, we additionally consider a deduplication-based preprocessing strategy. Specifically, we compute the normalized ℓ_2 (NL2) distance between image embeddings and remove samples whose distance is smaller than 1.4. This threshold has been commonly used in prior work as a practical criterion for identifying near-duplicate images Chen et al. (2024b). It is important to note that this deduplication procedure is not perfect in our setting. Due to the random data augmentations applied during training, two images originating from the same source may undergo different stochastic transformations and therefore may not always be detected as duplicates. Consequently, some visually similar samples may remain in the training set after deduplication. We intentionally adopt this imperfect deduplication strategy to better reflect real-world data collection

	Memorization			Generation Quality			
Method	SSCD↓	NL2↑	CLIP Emb.↓	CLIP↑	AES↑	LPIPS↑	IR↑
W/O	0.76	0.79	0.88	33.20	5.08	0.39	0.40
Prompt Optim	0.66	0.88	0.78	31.50	4.60	0.48	0.26
CFG-reverse	0.79	0.81	0.88	33.24	4.89	0.48	0.17
RTA	0.72	0.95	0.74	33.39	5.10	0.52	0.15
BAF-Soft	0.57	1.11	0.64	33.40	5.43	0.72	0.53
BAF-Hard	0.54	1.13	0.61	32.38	5.02	0.72	0.50

Table 4: Results on the **Pokémon** dataset using Stable Diffusion 1.5. Lower SSCD and CLIP embedding similarity indicate reduced memorization, while higher NL2 indicates larger distance from memorized training samples. Generation quality is evaluated using CLIP score, AES, LPIPS, and ImageReward (IR).

processes. In practice, training datasets often contain images that are highly similar but not strictly identical, for example due to cropping, color variations, or other preprocessing operations.

In our experiments, memorization strength is simulated by duplicating a set of canary images. If strict deduplication were applied to remove exact copies, this setting would no longer capture the realistic scenario where near-duplicate images appear in the training set. Our NL2-based filtering therefore provides a more faithful approximation of real-world data redundancy while still reducing trivial duplicates.

D.1 Experimental results on additional metrics

Due to space limitations in the main paper, we only report results on three backbones and three representative metrics: SSCD, CLIP score, and AES. To provide a more comprehensive evaluation, we report additional metrics on Pokemon dataset and celebA-HQ dataset and organize them into two groups: *memorization* and *generation quality* and one additional model SD 2.1.

Memorization metrics. To better quantify memorization behavior, we additionally report: (1) **CLIP embedding similarity**, which measures the cosine similarity between the generated image and the corresponding canary image in the CLIP embedding space; and (2) **Normalized ℓ_2 distance (NL2)**, which measures the pixel-level distance between generated images and canary images after normalization. Lower CLIP embedding similarity and higher NL2 indicate weaker memorization.

Generation quality metrics. To evaluate generation quality beyond CLIP score and AES, we further report: (1) **ImageReward (IR)**, a learned reward model reflecting human preference for text-to-image alignment; and (2) **LPIPS**, a perceptual similarity metric widely used to measure visual diversity and perceptual distance between images. Higher ImageReward and LPIPS generally indicate better generation quality and diversity.

Due to computational budget constraints, we report these additional metrics for Stable Diffusion 1.5, Stable Diffusion 2.1, and SDXL. The corresponding evaluation on FLUX is left for future work. As shown in the Table 3 to the Table 8, across these additional metrics, our method consistently mitigates memorization while maintaining or even improving generation quality.

E Additional Qualitative Results and Human Evaluation

To further evaluate BAF under realistic deployment settings, we provide additional qualitative comparisons on four community-released LoRA adapters.

Unlike the representative examples shown in the main paper, the results presented here are sampled directly from our evaluation set without any manual cherry-picking. For each LoRA, all compared methods use identical prompts, random seeds, and generation settings. Since all evaluated adapters are downloaded directly from community repositories, the degree of memorization naturally varies across LoRAs. Consequently, not every example exhibits dramatic visual differences, reflecting the diversity of memorization encountered in real-world LoRA releases. In particular, the LoRAs shown

	Memorization			Generation Quality			
Method	SSCD↓	NL2↑	CLIP Emb.↓	CLIP↑	AES↑	LPIPS↑	IR↑
W/O	0.74	0.83	0.86	33.16	4.87	0.40	0.45
Prompt Optim	0.71	0.88	0.78	29.80	4.60	0.49	0.27
CFG-reverse	0.78	0.84	0.75	33.39	4.89	0.50	0.14
RTA	0.68	0.92	0.74	30.02	5.10	0.54	0.16
BAF-Soft	0.62	1.09	0.64	34.27	5.71	0.74	0.59
BAF-Hard	0.60	1.12	0.61	34.23	5.49	0.71	0.55

Table 5: Results on the **Pokémon** dataset using Stable Diffusion XL. Lower SSCD and CLIP embedding similarity indicate reduced memorization, while higher NL2 indicates larger distance from memorized training samples. Generation quality is evaluated using CLIP score, AES, LPIPS, and ImageReward (IR).

	Memorization			Generation Quality			
Method	SSCD↓	NL2↑	CLIP Emb.↓	CLIP↑	AES↑	LPIPS↑	IR↑
W/O	0.77	0.83	0.86	31.80	5.00	0.40	0.42
Prompt Optim	0.69	0.87	0.78	29.40	4.58	0.49	0.27
CFG-reverse	0.73	0.81	0.85	30.60	4.88	0.49	0.12
RTA	0.70	0.89	0.75	29.80	5.08	0.53	0.16
BAF-Soft	0.57	1.08	0.67	33.50	5.58	0.72	0.55
BAF-Hard	0.55	1.07	0.62	32.80	5.24	0.74	0.52

Table 6: Results on the **Pokémon** dataset using Stable Diffusion 2.1. Lower SSCD and CLIP embedding similarity indicate reduced memorization, while higher NL2 indicates larger distance from memorized training samples. Generation quality is evaluated using CLIP score, AES, LPIPS, and ImageReward (IR).

in Figures 4 and 6 exhibit relatively weak memorization, whereas those in Figures 5 and 7 contain substantially stronger memorized content.

E.1 Blind Human Evaluation

To quantitatively evaluate the qualitative results, we conduct a blind human evaluation on **20** community-released LoRA adapters. The evaluation includes the four representative qualitative comparisons shown in Figures 4– 7 together with sixteen additional LoRAs omitted from the appendix due to space limitations. Each qualitative comparison consists of high-resolution images for detailed visual inspection and therefore occupies a substantial amount of space. For this reason, only four representative comparisons are presented in the appendix, while all **20** LoRAs are included in the user study and the reported statistics.

Because real-world LoRAs exhibit varying degrees of memorization, we first categorize all LoRAs into *high-memorization* and *low-memorization* groups before conducting the user study. The categorization is based solely on the visual similarity between images generated by the original LoRA and its representative subject. LoRAs whose generations consistently reproduce a recognizable identity or highly similar visual appearance are categorized as *high memorization*, whereas those without obvious identity reproduction are categorized as *low memorization*. This categorization is used only to determine the evaluation protocol and does not affect the comparison among methods.

Each participant evaluates all 20 LoRAs independently, and the display order of the compared methods is randomized for every qualitative comparison.

Low-memorization LoRAs (e.g., Figures 4 and 6). Since these LoRAs exhibit little observable memorization, participants evaluate only style preservation.

Q1 (Style Preservation).

	Memorization			Generation Quality			
Method	SSCD↓	NL2↑	CLIP Emb.↓	CLIP↑	AES↑	LPIPS↑	IR↑
W/O	0.93	0.92	0.89	29.74	5.46	0.42	0.78
Prompt Optim	0.68	0.81	0.78	28.85	5.10	0.50	0.73
CFG-reverse	0.71	0.84	0.84	29.65	5.53	0.47	0.73
RTA	0.74	0.79	0.79	29.11	5.34	0.52	0.67
BAF-Soft	0.52	1.42	0.60	31.45	5.67	0.63	0.72
BAF-Hard	0.44	1.42	0.57	31.24	5.51	0.65	0.71

Table 7: Results on the **CelebA-HQ** dataset using Stable Diffusion 1.5. Lower SSCD and CLIP embedding similarity indicate reduced memorization, while higher NL2 indicates larger distance from memorized training samples. Generation quality is evaluated using CLIP score, AES, LPIPS, and ImageReward (IR).

	Memorization			Generation Quality			
Method	SSCD↓	NL2↑	CLIP Emb.↓	CLIP↑	AES↑	LPIPS↑	IR↑
W/O	0.84	0.96	0.88	29.03	4.72	0.47	0.75
Prompt Optim	0.74	1.07	0.79	27.90	4.70	0.47	0.70
CFG-reverse	0.75	1.04	0.82	30.02	5.28	0.53	0.71
RTA	0.81	0.99	0.85	28.27	4.91	0.49	0.66
BAF-Soft	0.68	1.23	0.69	32.45	6.13	0.61	0.76
BAF-Hard	0.67	1.21	0.67	32.36	6.10	0.61	0.74

Table 8: Results on the **CelebA-HQ** dataset using Stable Diffusion XL. Lower SSCD and CLIP embedding similarity indicate reduced memorization, while higher NL2 indicates larger distance from memorized training samples. Generation quality is evaluated using CLIP score, AES, LPIPS, and ImageReward (IR).

Please rate how well the generated image preserves the artistic style of the original LoRA on a five-point Likert scale:

- 5: Perfectly preserves the original artistic style.
- 4: Largely preserves the original artistic style.
- 3: Moderately preserves the original artistic style.
- 2: Preserves only a small portion of the original artistic style.
- 1: Fails to preserve the original artistic style.

High-memorization LoRAs (e.g., Figures 5 and 7). Since these LoRAs exhibit recognizable memorized identities, participants evaluate both style preservation (Q1) and memorization removal (Q2).

Q2 (Memorization Removal).

Please rate how effectively the generated image removes memorized content while remaining faithful to the prompt on a five-point Likert scale:

- 5: No recognizable memorized identity remains.
- 4: Most memorized characteristics are removed.
- 3: Memorized characteristics are partially removed.
- 2: Only a small amount of memorization is removed.
- 1: The generated image remains highly similar to the memorized subject.

Table 10 summarizes the average ratings from ten participants over all 20 evaluated LoRAs. Across all evaluation criteria, BAF achieves the highest average score, indicating that it consistently suppresses

Method	Memorization			Generation Quality			
	SSCD↓	NL2↑	CLIP Emb.↓	CLIP↑	AES↑	LPIPS↑	IR↑
W/O	0.88	0.94	0.90	29.40	5.08	0.50	0.76
Prompt Optim	0.71	1.03	0.81	28.30	4.90	0.49	0.72
CFG-reverse	0.73	1.00	0.81	29.84	5.40	0.54	0.73
RTA	0.77	0.98	0.83	28.72	5.12	0.51	0.65
BAF-Soft	0.60	1.30	0.64	31.91	5.90	0.59	0.74
BAF-Hard	0.57	1.32	0.62	31.79	5.82	0.58	0.73

Table 9: Results on the **CelebA-HQ** dataset using Stable Diffusion 2.1. Lower SSCD and CLIP embedding similarity indicate reduced memorization, while higher NL2 indicates larger distance from memorized training samples. Generation quality is evaluated using CLIP score, AES, LPIPS, and ImageReward (IR). Best results are highlighted in bold.



Figure 4: Additional qualitative comparison on the *Paper Artistry* LoRA. This example is included directly from the evaluation set without manual selection. Although the original LoRA exhibits only moderate memorization, BAF better preserves the paper-art style while avoiding recognizable memorized content.

memorized content while preserving the artistic style learned by the original LoRA. These results further support the qualitative observations presented in Figures 4– 7.



Figure 5: Additional qualitative comparison on the *Clay Vorona* LoRA. This LoRA exhibits relatively stronger memorization, making the differences among methods more pronounced. BAF suppresses memorized content while preserving the characteristic clay sculpture appearance more faithfully than the inference-time baselines.

Table 10: Blind human evaluation on 20 community-released LoRA adapters. Four representative examples are shown in Figures 4–7, while 16 additional examples are omitted due to space limitations. Ten participants were unaware of the corresponding methods. Higher scores indicate better performance.

Method	**Style ↑**	**De-Memo ↑**	**Overall ↑**
CFG	3.63	3.18	3.40
Prompt Optim.	3.81	3.47	3.71
RTA	4.09	3.91	4.00
BAF (Ours)	**4.76**	**4.68**	**4.72**

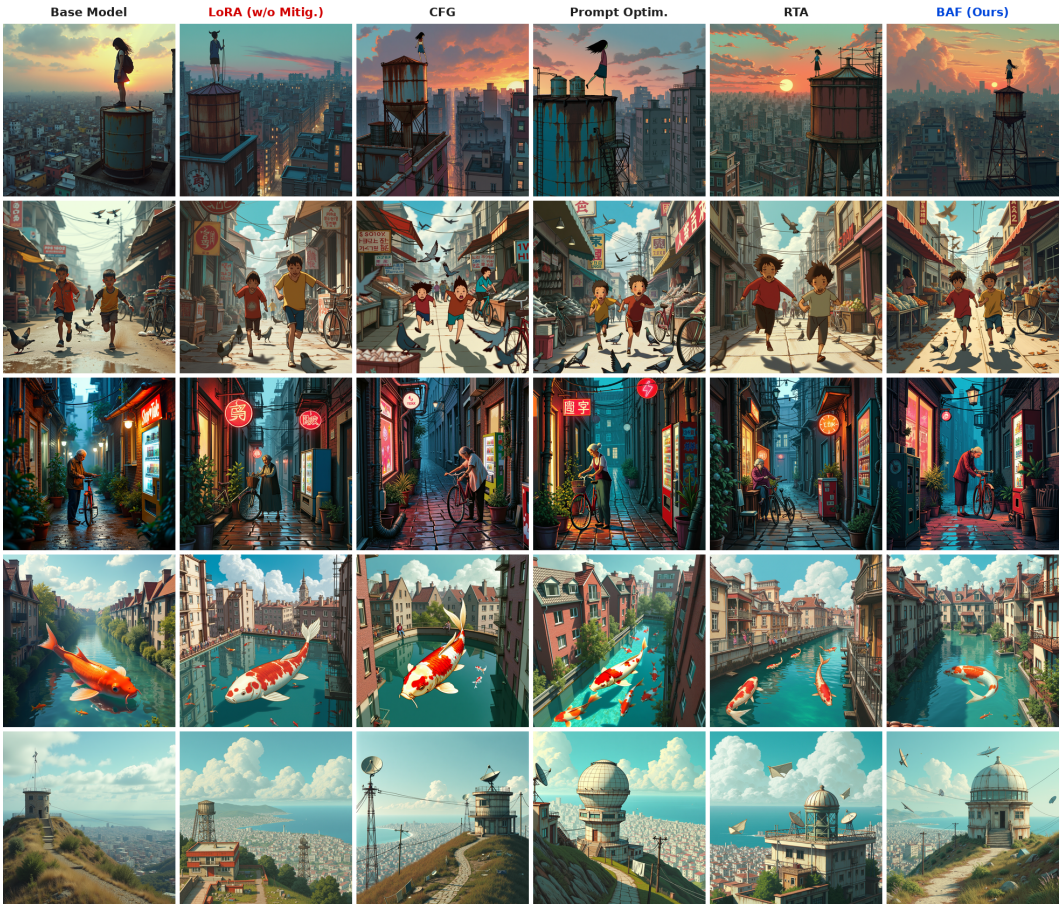

Figure 6: Additional qualitative comparison on the *Tekkonkinkreet Style* LoRA. Although the memorization in this LoRA is relatively weak, BAF maintains the original artistic style without introducing noticeable degradation, remaining competitive with or better than the inference-time baselines.



Figure 7: Additional qualitative comparison on the *Disney Nouveau* LoRA. BAF consistently preserves the distinctive illustration style while reducing recognizable memorized characteristics.

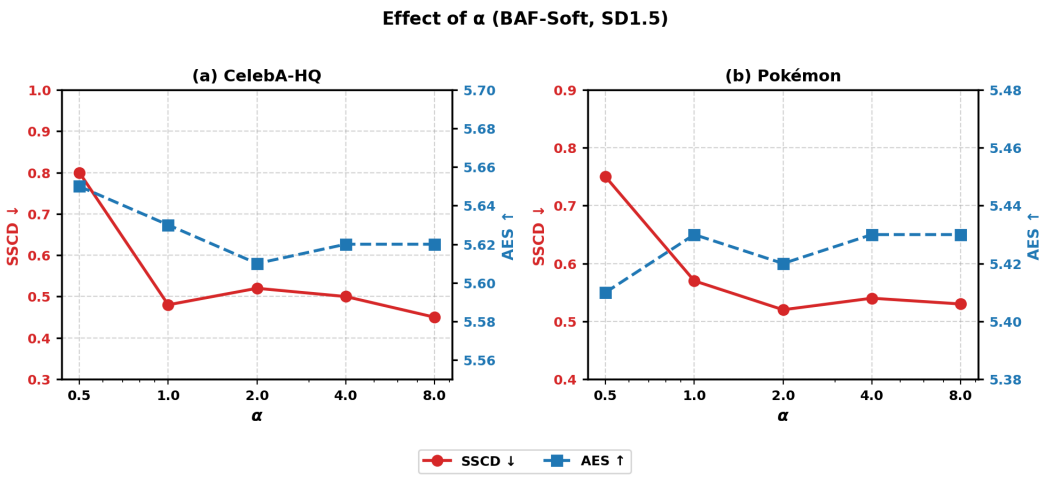

Figure 8: **Effect of α in BAF-Soft on CelebA-HQ (a) and Pokémon (b), both trained on SD1.5.** We measure SSCD (red) and AES (blue).

F Experiments about Additional Analyses

Effect of α in BAF-Soft. Figure 8 studies the effect of the scaling parameter α in BAF-Soft. Across both CelebA-HQ and Pokémon, the most notable drop in SSCD occurs when α increases from 0.5 to 1, indicating that even a moderate scaling factor is sufficient to substantially suppress memorization. Beyond $\alpha = 1$, SSCD generally decreases or stabilizes with only minor fluctuations, suggesting that the benefit of further increasing α diminishes. Importantly, the aesthetic score (AES) remains nearly constant across the entire range of α on both datasets, with a variation of less than 1%. These results suggest that α primarily controls the strength of memorization suppression, while having minimal impact on generation quality, and that $\alpha = 1$ offers a favorable trade-off between the two objectives.

Effect of τ_{energy} in BAF-Soft and visualization. Figure 9 examines how the energy threshold τ_{energy} affects memorization (SSCD) and generation quality (AES) on CelebA-HQ and Pokémon. When τ_{energy} is small, memorization is strongly suppressed but AES is slightly reduced, suggesting that some useful adaptation capacity is also weakened. As τ_{energy} increases, AES improves while SSCD increases moderately, indicating a better memorization–utility trade-off. However, when $\tau_{\text{energy}} = 1.0$, the filtering effect largely disappears, leading to a sharp increase in SSCD.

To understand the origin of this behavior, we analyze the distribution of anchoring scores under different thresholds in Figure 10. The key factor is that the energy threshold τ_{energy} determines the dimension K of the selected principal subspace of the pretrained model. As τ_{energy} increases, more singular directions are included, which enlarges K and therefore raises the null baseline $a_{\text{null}} = (K/m)(K/n)$.

When τ_{energy} is small, the selected principal subspace is very compact. In this regime, the anchoring scores concentrate within a narrow range near zero. Although BAF-Soft weakens many low-alignment directions, the compressed score range limits the ability to reliably distinguish memorized channels from generalizable ones. As τ_{energy} increases, the score distribution expands and becomes increasingly structured. At $\tau_{\text{energy}} = 0.85$, a clear valley emerges between low- and high-alignment regions. Importantly, the baseline a_{null} lies close to this valley, allowing BAF-Soft to suppress low-alignment (memorized) directions while preserving high-alignment components that contribute to generalizable adaptation. When $\tau_{\text{energy}} = 1.0$, the selected subspace spans nearly the entire spectral space of the pretrained weights. In this case, K approaches $\min(m, n)$ and a_{null} becomes close to one. Consequently, most channels obtain anchoring scores close to the upper end of the range, causing the soft gating function $g(a_i) = a_i^{\alpha}$ to approach 1 for the majority of channels. The filtering effect therefore largely disappears, reintroducing memorized directions and leading to the observed increase in SSCD. These observations suggest that overly small or overly large values of τ_{energy} both weaken

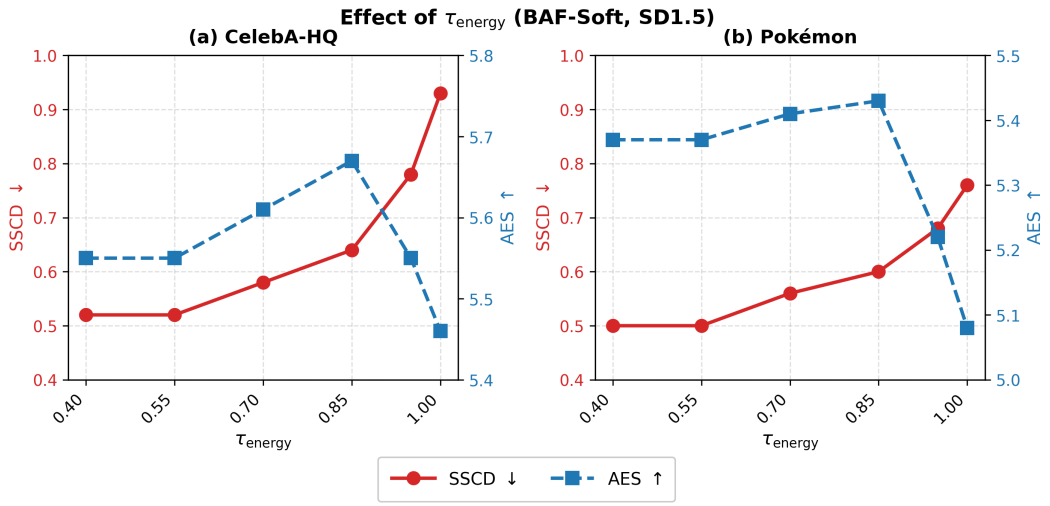

Figure 9: **Effect of τ_{energy} in BAF-Soft on CelebA-HQ (a) and Pokémon (b), both trained on SD1.5.** We measure SSCD (red) and AES (blue).

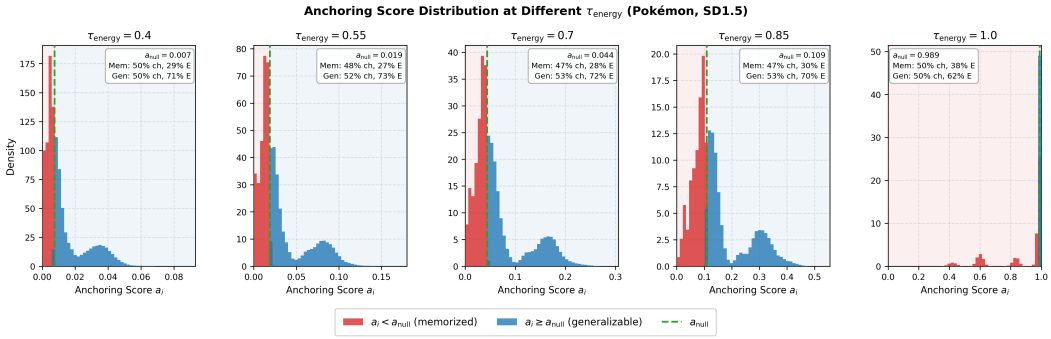

Figure 10: **Anchoring score distributions of the LoRA at different τ_{energy} (Pokémon, SD1.5).** Each panel shows the distribution of per-channel anchoring scores a_i for a given τ_{energy}. Red bars indicate channels below the null threshold a_{null} (memorized); blue bars indicate channels at or above it (generalizable). The green dashed line marks $a_{\text{null}} = (K/m)(K/n)$, where K is selected by the energy criterion. The inset reports the fraction of channels and total energy in each region.

the discriminative power of anchoring scores, while intermediate values provide a more effective separation between memorized and reusable directions.

Mitigation under Different Memorization Strengths. To systematically study memorization behavior, we simulate different memorization strengths by duplicating training images during LoRA fine-tuning. Specifically, each canary image is repeated D times in the training set, where D denotes the duplication factor. Larger D therefore induces stronger memorization pressure.

To measure memorization under different regimes, we report two complementary metrics: *targeted SSCD* and *Top-5% SSCD*. Targeted SSCD measures similarity between generated images and a specific duplicated training instance, capturing instance-level memorization of the canary sample. However, when the duplication factor D is small, memorization may not concentrate on a particular instance and it becomes unclear which training image is memorized. In this regime, targeted SSCD becomes less informative. Hence, we also report *Top-5% SSCD*, which measures the 95th percentile similarity between generated images and the training dataset. This metric is commonly used to detect dataset-level memorization by identifying whether a small fraction of generated samples are unusually close to training images.

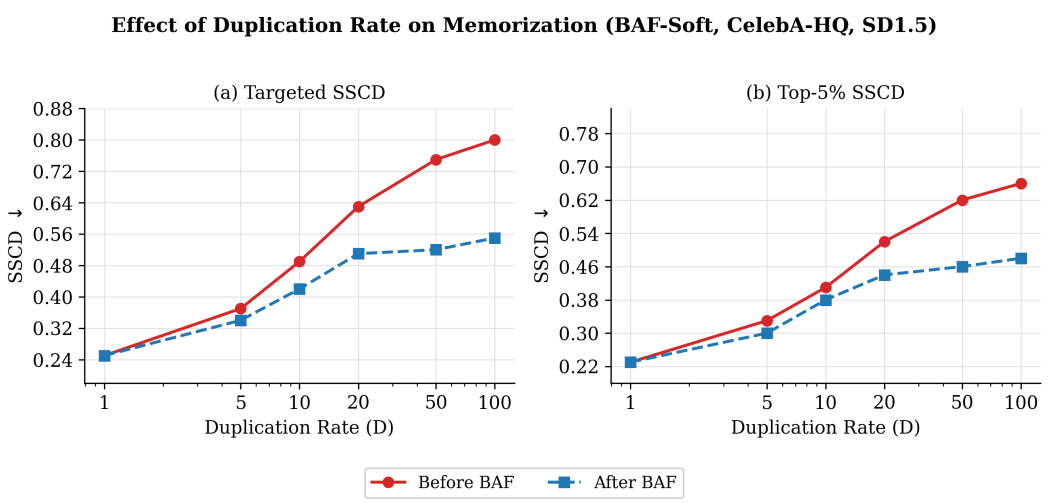

Figure 11: Effect of duplication rate on memorization during LoRA fine-tuning. To simulate different memorization strengths, each canary image is duplicated D times during training. The left panel reports *targeted SSCD* and the right panel reports *Top-5% SSCD* Red curves denote the model before applying BAF, while blue curves denote the results after applying BAF.

Table 11: Runtime of the one-time BAF filtering procedure on a single NVIDIA A100 GPU.

Backbone	Base Keys	Matched Layers	Layer Size	Time (s)
SD1.5	686	128	320×320	38.6
SDXL	1680	560	640×640	343.3

Figure 11 shows that after applying BAF (blue curves), memorization is consistently reduced across all duplication levels. Importantly, even when duplication is high (e.g., $D = 50$ or 100), the resulting SSCD values remain around 0.5, which prior work commonly considers to indicate a low likelihood of memorization. These results demonstrate that BAF effectively suppresses memorization signals across a wide range of memorization strengths.

Qualitative Analysis under Different Hyperparameters Beyond the quantitative analyses, we further visualize how different hyperparameter choices affect the generated images. Figures 12 and 13 present representative examples under different values of τ_{energy} and α, respectively.

Figure 12 illustrates the effect of varying τ_{energy} while fixing $\alpha = 0.4$. When τ_{energy} is relatively small, BAF performs stronger filtering, effectively suppressing memorized characteristics while preserving most of the learned style. As τ_{energy} increases, more principal directions are retained, causing the generated images to gradually approach those of the original LoRA. Consequently, stylistic details become increasingly faithful to the original adapter, but recognizable memorized characteristics also begin to reappear. This continuous visual transition is consistent with the quantitative trend shown in Figure 9.

Figure 13 studies the influence of the soft-gating exponent while fixing $\tau_{\text{energy}} = 0.8$. Increasing α strengthens the suppression of low-anchoring channels, resulting in progressively stronger memorization mitigation. Moderate values (around $\alpha = 0.4$–0.5) preserve the artistic style while effectively reducing memorization. When α becomes excessively large, however, useful adaptation directions are also attenuated, leading to noticeable semantic changes and degradation of style fidelity. Overall, the qualitative observations demonstrate that τ_{energy} and α provide intuitive control over the trade-off between memorization suppression and style preservation.

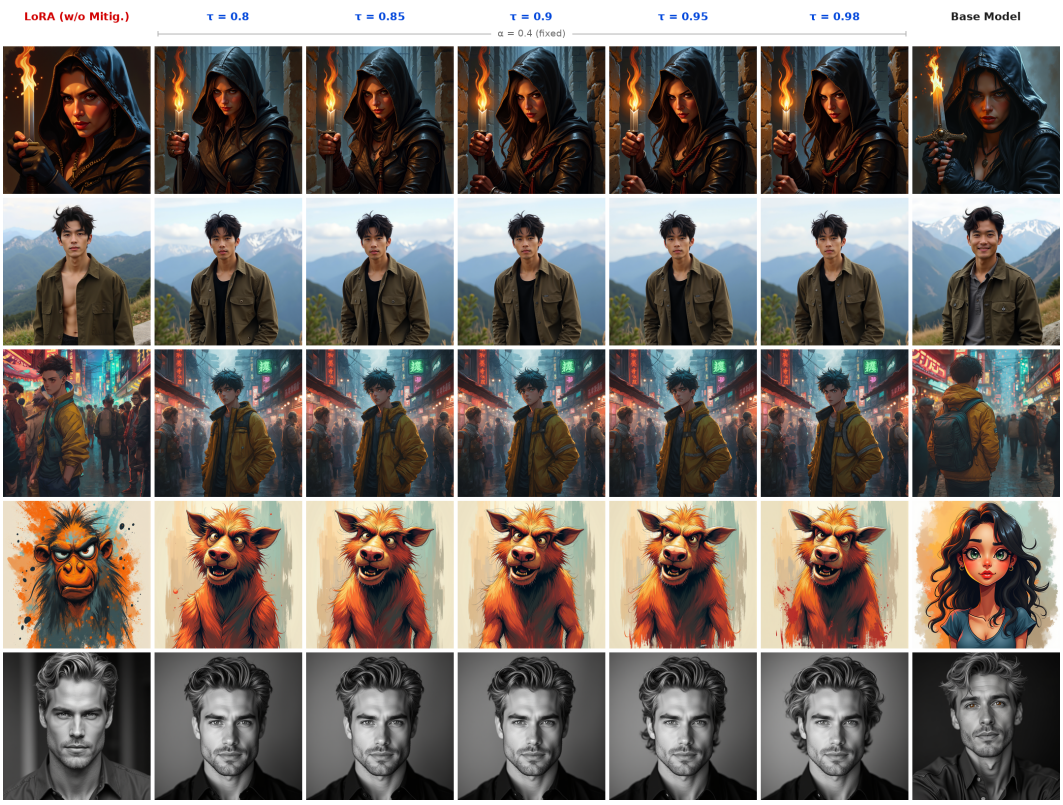

Figure 12: **Qualitative effect of varying the energy threshold τ_{energy} while fixing $\alpha = 0.4$.** Each row corresponds to a different prompt. As τ_{energy} increases, more principal directions of the pretrained backbone are retained, causing the filtered LoRA to gradually approach the original LoRA. Moderate values provide a favorable balance between suppressing memorized characteristics and preserving the learned artistic style.

G Runtime Analysis

Table 11 reports the runtime of BAF on a single NVIDIA A100 GPU for two representative diffusion backbones, Stable Diffusion 1.5 (SD1.5) and SDXL.

BAF consists of three computational stages: (1) computing the principal subspace of the pretrained backbone through singular value decomposition (SVD), (2) computing the thin SVD of the LoRA weight update, and (3) evaluating anchoring scores followed by channel-wise filtering. Among these stages, the SVD of the pretrained backbone dominates the overall runtime, whereas the remaining operations are comparatively lightweight.

Importantly, BAF is an entirely *offline* post-processing method. The filtering procedure is executed only once before releasing the LoRA adapter, after which the filtered weights are stored and used directly during inference. Consequently, BAF introduces *no additional inference-time overhead*, in contrast to inference-time mitigation methods that require extra computation for every generated image.

Furthermore, the dominant computation depends only on the pretrained backbone rather than the released LoRA. For a fixed backbone, the corresponding principal subspace can therefore be precomputed once and reused across all LoRA adapters trained on the same base model. Under this caching strategy, filtering additional LoRAs only requires processing the lightweight LoRA updates, substantially reducing the amortized runtime in practical deployment.

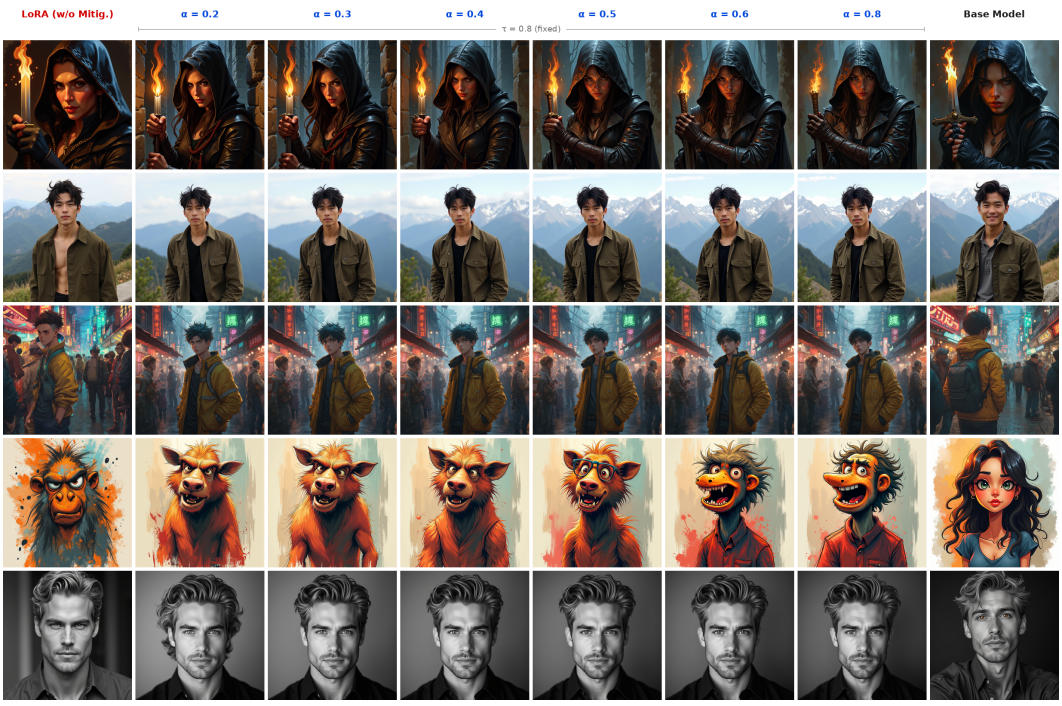

Figure 13: **Qualitative effect of varying the soft-gating exponent α while fixing $\tau_{\text{energy}} = 0.8$.** Larger values of α impose stronger attenuation on low-anchoring channels. Moderate values effectively suppress memorized content while preserving the artistic style of the original LoRA, whereas excessively large values begin to remove useful adaptation, resulting in visible degradation of style fidelity.