AUTOANCHOR: STABLE DIFFUSION UNLEARNING USING CROSS-ATTENTION AS A MANIFOLD SURROGATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion unlearning is essential for mitigating the generation of harmful or copyrighted content in text-to-image models. Current diffusion unlearning techniques determine the model update direction by either using alternatives of the target concept as anchors or using empty prompts. The anchor-based method relies on manually and semantically-chosen anchors that risk biased unlearning, while the anchor-free method inherently suffers from imprecise unlearning due to unconstrained latent updates. In this work, we theoretically formalize diffusion unlearning under the manifold hypothesis and prove that lacking a manifold-proximal anchor can induce substantial normal-space drift that degrades unlearning performance. Motivated by this analysis, we propose AutoAnchor, a two-stage framework that automatically synthesizes manifold-proximal anchors. However, direct geometric manifold optimization is computationally intractable. To address this challenge, AutoAnchor introduces a novel cross-attention consistency loss, which serves as a highly efficient surrogate for manifold proximity. Experimental results demonstrate that AutoAnchor effectively achieves precise and unbiased unlearning across various state-of-the-art baselines, significantly improving targeted concept removal (by up to 28.17% in CLIP score) and non-target utility (by up to 4.17% in CLIP score). Moreover, AutoAnchor can also be easily integrated into existing diffusion unlearning methods to enhance their unlearning performance (by 34.48% for concept removal and 32.63% for utility on average).

1 INTRODUCTION

Text-to-Image diffusion models Ho et al. (2020) have achieved remarkable success in generating high-fidelity images conditioned on input text prompts. However, their reliance on massive and uncurated web data introduces significant risks, *e.g.*, generating harmful content or copyright infringement. To mitigate these risks, diffusion unlearning has emerged as a crucial technique. The goal of diffusion unlearning is to erase a diffusion model's ability to generate specific target concepts, which is usually realized by tuning the model parameters towards certain update directions.

Achieving effective erasure while preserving model utility is difficult for modern diffusion models. This is because they operate in a high-dimensional latent space, but the valid and realistic images reside on a low-dimensional data manifold within that. Without a carefully designed update direction, diffusion unlearning can easily push the model off the data manifold when erasing target concepts, thereby damaging the underlying geometry and risking the model's utility degradation.

Regarding how to determine the update directions, current diffusion unlearning methods are categorized into *anchor-free* approaches and *anchor-based* approaches. *Anchor-free* methods (*e.g.,* Gandikota et al. (2023); Han et al. (2024); Li et al. (2025a)) use an empty or generic prompt as directions for unlearning, simply pushing the generation away from the target concepts. However, the resulting update directions need not align with the local structure around the target concept, creating a off-manifold drift risk. This geometric damage can lead to **imprecise unlearning**, manifesting as catastrophic forgetting Heng & Soh (2023), degraded image quality Fan et al. (2023); Lee et al. (2025a), and weakened text-image alignment Ko et al. (2024); Han et al. (2024) on non-target concepts. Moreover, such methods risk fragile forgetting that can be easily bypassed by adversarial prompts Bui et al. (2024); Wu et al. (2025b); Zhang et al. (2024b) or simple fine-tuning Pan et al. (2024). Existing *anchor-based* methods (*e.g.,* Kumari et al. (2023); Lyu et al. (2024)) provide an

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.

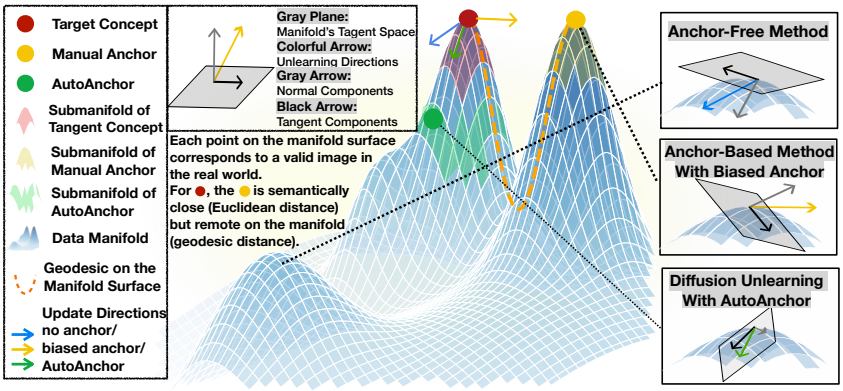

Figure 1: Illustration of the data manifold and unlearning update directions. Without a manifold-proximal anchor (*e.g.*, our AutoAnchor), the update direction can risk normal drift that pushes the model off the manifold, leading to imprecise unlearning (ineffective removal or utility degradation).

explicit alternative concept (the anchor). However, the anchors that they heavily rely on are manually chosen based on semantic similarity, which risks geometry mismatch because an anchor can be semantically similar but manifold-distinct. With such biased anchors, the unlearning can still suffer from the imprecise unlearning issues above and may further cause non-robust forgetting Wu et al. (2025a).

Both approaches overlook the need to maintain consistency with the underlying manifold structure during the unlearning process. To fill this gap, we analyze the unlearning process under the manifold hypothesis. Geometrically, valid images lie on a low-dimensional data manifold within the vast diffusion latent space, as illustrated in Figure 1. A robust and effective unlearning update should smoothly traverse along this manifold to alter target concepts, rather than pushing the representation away from it. Therefore, we decompose the unlearning update directions into their tangential (along the manifold) and normal (orthogonal to the manifold) components.

For anchor-free methods, the empty prompt cannot explicitly prevent off-manifold update; for manual-anchor methods, semantic similarity alone does not guarantee tangent update as well. Our analysis quantifies a lower bound of their normal drift, providing a potential explanation why these unlearning can be imprecise. We further show how manifold-proximal anchors suppress the normal drift and why they are useful to realize unbaised and precise diffusion unlearning.

However, constructing such a manifold-proximal anchor presents non-trivial challenges. Direct optimization on the data manifold is computationally intractable Chung et al. (2022); Huguet et al. (2022); Park et al. (2023), as the manifold is only implicitly defined by the complex distribution of the diffusion model's latent space. Furthermore, there exists a fundamental domain gap: while the unlearning update operates in the diffusion latent space, the anchor is optimized within the vast prompt embedding space. To bridge this gap, our key insight is that in diffusion architectures, a frozen model's cross-attention maps explicitly dictate the layout of certain concepts when generating images, which implicitly reflect the data manifold learned by the model. By leveraging additional semantic constraints, these attention maps serve as a powerful and efficient proxy for local manifold alignment, with a conditional normal-energy bound derived in Appendix A.7. This allows us to optimize manifold-proximal anchors by aligning their attention patterns with the target concept.

To this end, we propose AutoAnchor, an efficient two-stage process that automatically synthesizes manifold-proximal anchors for diffusion unlearning. It first filters and aggregates valid candidate concepts automatically, and then optimizes the anchors to be manifold-proximal to the target concept. We further propose the cross-attention consistency loss, a novel manifold optimization surrogate specific to anchor construction. AutoAnchor empirically mitigates off-manifold drift and thus effectively improves the removal-utility trade-off. Moreover, it can be integrated into related works to further improve their unlearning performance. Our main contributions are summarized as follows:

- **Theoretical Formalization of Diffusion Unlearning Issues:** We analyze the underlying geometry of diffusion unlearning, identifying that anchor-free methods risk manifold shift and manually-

anchor methods risk manifold misalignment. This gives a geometric insight of anchor sensitivity and motivates controlling the normal drift when constructing an anchor for diffusion unlearning.

- **Novel Manifold-Proximal Surrogate and Unlearning Framework:** To overcome the computational intractability and domain gap of direct manifold optimization, we introduce a cross-attention consistency loss. Using this surrogate loss, our proposed AutoAnchor efficiently and automatically constructs manifold-proximal anchors for realizing unbiased and precise diffusion unlearning. Furthermore, we theoretically explain how AutoAnchor approximates manifold proximity.
- **Empirical Evaluation:** Extensive experiments show that AutoAnchor improves the diffusion unlearning performance of the state-of-the-art baselines in both concept removal (additionally decrease 34.48% CLIP score on average) and model utility (reduce 32.63% of CLIP degradation on average). Moreover, ablation studies show the effects of cross-attention loss, and multiple unlearning shows the robustness of AutoAnchor.

2 Related Works

Based on how they determine the update direction, existing diffusion unlearning methods are categorized into anchor-free and anchor-based approaches. **Anchor-free** approaches determine their update directions using coarse signals like empty prompts, generic concepts, or unconditional scores. For example, ESD Gandikota et al. (2023) utilizes the difference of predictions between the target concept and empty prompt as the update direction. Such imprecise directions usually lead to imprecise unlearning with an undesirable removal–utility trade-off. Although existing works have made efforts to address them, they target individual issues in isolation. Specifically, Zhang et al. (2024b) only focuses on robustness against adversarial prompts; Ko et al. (2024); Heng & Soh (2023); Shirkavand et al. (2025); Amara et al. (2025); Fan et al. (2023) only aim to strengthen model utility; Lu et al. (2024) is solely concerned with utility loss after multiple unlearning; and Pan et al. (2024); George et al. (2025); Xiong et al. (2025) only pursue effective concept removal. However, with the imprecise update directions, it is difficult to solve these issues simultaneously. Our method aims to solve this by using a manifold-proximal anchor, which will provide a manifold-coherent update direction theoretically. With less normal drift during unlearning, AutoAnchor empirically achieves more precise removal and better non-target utility preservation.

Anchor-based approaches pre-define specific anchor concepts for the target concepts, aiming to provide more precise unlearning directions than anchor-free methods. For example, ConAbl Kumari et al. (2023) first uses the generation direction differences between the target and anchor concepts as the update direction, thus realizing better utility preservation than ESD. Recent anchor-based methods further explore more sophisticated problems like reliable erasure Hong et al. (2024); Wu et al. (2025b); Li et al. (2025a); Wu et al. (2025a), multiple unlearning Tu et al. (2026); Han et al. (2024), training-free unlearning Biswas et al. (2025), concept erosion Lyu et al. (2024); Park et al. (2024); Thakral et al. (2025); Lee et al. (2025a), and attention-based unlearning Zhang et al. (2024a; 2025a). However, existing anchor-based methods heavily rely on the assumption that their manually chosen anchors are precise. This assumption is not practical when semantically close anchors are not manifold-proximal to the target concepts, leading to biased unlearning as we show in Appendix §C.7. While some works Wu et al. (2025a); Lyu et al. (2024); Lee et al. (2026b); Kim et al. (2026) attempt to generate anchors automatically, they still rely on semantic closeness and semantic closeness alone does not explicitly control the geometry bias induced by unlearning. AutoAnchor solves this by adding cross-attention optimization step as a practical manifold-proximity surrogate.

3 Preliminaries

Latent Diffusion Models (LDMs). Modern text-to-image models (*e.g.*, Stable Diffusion Rombach et al. (2022)) operate in a compressed latent space. Given an input text prompt, the model first maps it into the corresponding embedding $\mathbf{e} \in \mathcal{E}$. Next, the model predicts the latent noise $\epsilon \in \mathbb{R}^d$, which is then used to generate an image latent $\mathbf{z} \in \mathbb{R}^d$ as the output. Training LDMs involves two processes. First, in the noise diffusion process, Gaussian noise ϵ is progressively added to the latent state $\mathbf{z}_0$ of a valid image, producing a noisy latent $\mathbf{z}_t$ at timestep $t \sim \mathcal{U}[0, T]$, where $\mathcal{U}$ denotes the uniform distribution. Second, in the denoising process, the diffusion model, a neural network parameterized by $\boldsymbol{\epsilon}_\theta : \mathbb{R}^d \times \mathcal{E} \times [0, T] \to \mathbb{R}^d$, predicts the added noise and learns the image manifold by minimizing

the standard objective $\mathcal{L}_{\mathrm{SD}}(\theta) = \mathbb{E}_{\mathbf{z}_0,\mathbf{e},t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}, t)\|^2\right]$. LDMs use the predicted noises to generate an image $\mathbf{z}_0$ by progressively denoising $\mathbf{z}_{\mathrm{t}-1} = \mathbf{z}_{\mathrm{t}} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}, t)$, for $t = T, ..., 2, 1$.

Diffusion Unlearning. Denote a pre-trained diffusion model $\boldsymbol{\epsilon}_{\theta^*}$ and a target concept $c_u \in \mathcal{C}$ to unlearn (e.g., "Mickey Mouse"). The goal of diffusion unlearning is to obtain a tuned model $\boldsymbol{\epsilon}_{\theta^u}$ that is incapable of generating images related to the target concept c_u when prompted, while preserving its generation capabilities for all non-target concepts. Existing unlearning algorithms achieve this by tuning the model's parameters along a specific update direction to alter its prediction for c_u. To formulate this update direction, we recall the known fact Ho et al. (2020); Song et al. (2021) that diffusion models learn the score function (*i.e.*, gradient of the log probability) of the data distribution, *i.e.*, $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}, t) \propto \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t|\mathbf{e})$. This equivalence allows us to represent unlearning updates (on the data manifold) as noise prediction differences (in latent space) under two distinct embeddings. For *anchor-free* (Δ^{AF}) and *anchor-based* (Δ^{AB}) methods, the update direction is expressed as:

$$\Delta^{\mathrm{AF}} = \boldsymbol{\epsilon}_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \boldsymbol{\epsilon}_{\theta^*}(\mathbf{z}_t, \mathbf{e}_\emptyset, t), \quad \Delta^{\mathrm{AB}} = \boldsymbol{\epsilon}_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \boldsymbol{\epsilon}_{\theta^*}(\mathbf{z}_t, \mathbf{e}_a, t), \tag{1}$$

where $\mathbf{e}_u$, $\mathbf{e}_a$, and $\mathbf{e}_\emptyset$ denote the embeddings for the target concept c_u, the anchor $c_a \in \mathcal{C}$, and the empty prompt, respectively. The geometric properties of the conditioning embedding (empty or anchor) critically determine unlearning stability, as we will formally analyze in §4.

4 Theoretical Analysis of Unstable Diffusion Unlearning

In §4.1, we first introduce the manifold notations and statistical assumptions. Next we theoretically analyze why existing diffusion unlearning methods can be unstable in §4.2. In §4.3, we further prove that unlearning updates induced by manifold-proximal anchors can suppress normal drift. Complete proofs, detailed assumption justifications, and extra stability analyses are in Appendix A.

4.1 Problem Setup and Statistical Assumptions

We consider a smooth and compact d^*-dimensional manifold embedded in the diffusion latent space $\mathcal{M} \subset \mathbb{R}^d$ which supports valid image latents. All data points associated with the target concept c_u support a d_u^*-dimensional submanifold $\mathcal{M}_u \subset \mathcal{M}$. To analyze the stability of unlearning updates, we decompose the prediction-difference directions relative to the data manifold $\mathcal{M}$. For any latent state $\mathbf{z}_u \in \mathcal{M}_u$, $T_{\mathcal{M}}$ denotes the d^*-dimensional *tangent space* (directions along the manifold surface), while $N_{\mathcal{M}} = T_{\mathcal{M}}^{\perp}$ denotes the $(d - d^*)$-dimensional *normal space* (directions orthogonal to the manifold). Fundamentally, effective unlearning requires erasing the target concept while maintaining the generation on the manifold to preserve utility. To preserve local structure, a desired update Δ should closely align with the tangent space, *i.e.*, $\Pi_{N_{\mathcal{M}}}\Delta \approx 0$ where Π denotes projection operation and $N_{\mathcal{M}}$ is the normal space for $\mathcal{M}$.[1]

Next we will introduce modeling assumptions on data distribution and LDMs. Assumption 1 describes the dimensional regime, Assumption 2 bounds prediction errors, and Assumption 3 models unlearning directions. They are all motivated by existing literature Cai et al. (2013); Pidstrigach (2022); Stanczuk et al. (2024); Li et al. (2025b); Chen et al. (2023); Vershynin (2018); Bubeck & Sellke (2021).

Assumption 1 (High-Dimensional Regime)**.** *The dimension of ambient latent space $\mathbb{R}^d$ is much larger than that of the data manifold $\mathcal{M}$ and target manifold $\mathcal{M}_u$, i.e., $d^*/d \to 0$ and hence $d_u^*/d \to 0$.*

Assumption 2 (Score Approximation Error)**.** *The noise-parameterized score approximation error $\varepsilon(\mathbf{z}_t, t) = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}, t) + \sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t|\mathbf{e})$ has zero mean, and its variance along each unit normal direction in $N_{\mathcal{M}}$ is bounded by σ_N^2, where $\sigma_t > 0$ is the forward noise standard deviation.*

Assumption 3 (Directional Isotropy)**.** *For the nonzero anchor-free and misaligned manual-anchor prediction differences analyzed below, we model their normalized directions $v_{AF} = \Delta^{AF}/\|\Delta^{AF}\|$ and $v_{AB} = \Delta^{AB}/\|\Delta^{AB}\|$ as uniform on the unit sphere conditional on the tangent space $T_{\mathcal{M}}$.*

We elaborate on the justifications and details of these assumptions in Appendix A.2 and A.8.2.

[1]The tangent-space relation and finite-step extension are given in Appendices A.8.1 and A.8.9.

4.2 Normal Drift in Existing Diffusion Unlearning Methods

Under Assumptions 1–3, we show the normal drift induced by anchor-free updates and geometrically misaligned manual anchors, which quantifies how much existing methods can be off-manifold.

Theorem 1 (Normal Drift without Tangent Alignment)**.** *For anchor-free or manual-anchor unlearning, the fraction of the update energy projected into the normal direction is concentrated:*

$$\forall \delta \in [0,1), \quad \mathbb{P}\left(\frac{\|\Delta_N\|^2}{\|\Delta\|^2} > 1 - \frac{d^*}{d} - \delta\right) \geq 1 - \exp\left(-\frac{d\delta^2}{8}\right). \tag{2}$$

Equation 2 means that with a probability larger than or equal to $1-\exp(-d\delta^2/8)$, the fraction of update energy in the normal direction (*i.e.*, $\|\Delta_N\|^2/\|\Delta\|^2$) is at least $1-d^*/d-\delta$. Note that for small δ with very large d, $1-d^*/d-\delta$ means a large normal-energy fraction, and $1-\exp(-d\delta^2/8)$ is close to 1, which is a high probability. Therefore, following anchor-free or misaligned manual anchors, most unlearning updating energy will lie in the normal space, providing a geometric explanation for a potential source of unlearning instability. We provide the absolute magnitude bound for $\|\Delta_N\|$ in Appendix A.3 (Theorem 3), which further quantifies this drift under an explicit error model.

4.3 Stabilizing Unlearning via Manifold-Proximal Anchors

To establish a controlled normal draft, we first define a ζ-proximal anchor based on the update direction it induces. Over latent states related to the target concept, an anchor is a ζ-proximal anchor if its induced update direction is predominantly along the local tangent space on average (with no larger than ζ^2 expected fraction of update energy in the normal direction). That is, a smaller ζ means more manifold-proximal (see the formal Definition 1 in Appendix A.1). Therefore, we have:

Theorem 2 (Suppression of Normal Drift via Proximal Anchors)**.** *For diffusion unlearning with a ζ-proximal anchor as in Definition 1, the normal component $\Delta_N^{PA} = \Pi_N(\Delta^{AB})$ satisfies:*

$$\forall \delta > 0, \quad \mathbb{P}\left(\frac{\|\Delta_N^{PA}\|^2}{\|\Delta^{PA}\|^2} \geq \zeta^2 + \delta\right) \leq \frac{\zeta^2}{\zeta^2+\delta}, \text{ where } \Delta^{PA} = \Delta^{AB}. \tag{3}$$

Theorem 2 shows that when ζ^2 is small relative to δ, with a low probability (*i.e.*, no larger than $\zeta^2/(\zeta^2+\delta)$), the normal energy fraction is at least $\zeta^2+\delta$. This reveals that with a tightly bounded ζ-proximal anchor, the induced update has a small expected normal-energy fraction, indicating potential for locally stable unlearning. We also provide the absolute magnitude bound for $\|\Delta_N^{\text{PA}}\|$ in Appendix A.4 (Theorem 4), which also controls the normal magnitude when the total update is bounded. Moreover, as proven via the local random-walk analysis in Appendix A (Theorem 5), this suppression tightens an upper bound on cumulative normal displacement under the stated dynamics. In §5, we introduce AutoAnchor to construct a manifold-proximal anchor in diffusion unlearning.

5 Design of AutoAnchor

To realize unbiased and precise diffusion unlearning, the anchor construction for AutoAnchor must satisfy three key requirements. **(R1) Non-Synonymous:** It must not be a direct synonym of the target concept. For example, using "naked" as an anchor fails to unlearn the "nude" concept, because the two distributions largely overlap, leaving the target insufficiently suppressed. **(R2) Semantic Alternative:** It should be a valid semantic alternative related to the target concept. For example, using an irrelevant concept "Mickey Mouse" as an anchor to unlearn the "nude" concept gives a misaligned semantic reference. **(R3) Manifold-Proximal:** It should be manifold-proximal to the target concept to guide a stable unlearning direction, as demonstrated by our theoretical analysis in §4. In addition to the three requirements of the anchor above, the generation process itself must be **automatic** (to avoid manually introducing extra subjective bias and ensure scalability of unlearning) and **efficient** (acting as a lightweight preprocessing step).

To meet these criteria, we propose a two-stage framework, illustrated in Figure 2. The first stage clusters and aggregates candidate alternatives to obtain a semantically motivated initialization (§5.1). The second stage refines its structural alignment using a cross-attention surrogate and a distance regularizer (§5.2). Finally, we use the constructed anchor in the removal objective.

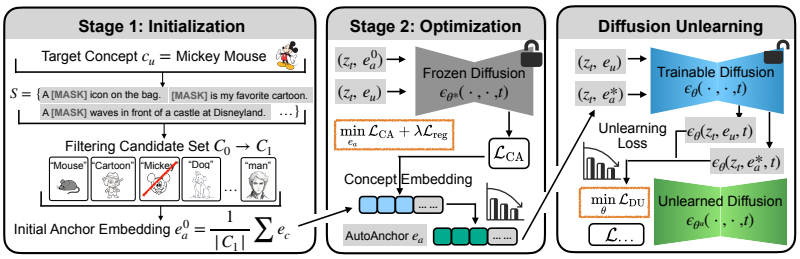

Figure 2: Illustration of AutoAnchor. Given a target concept c_u to unlearn, AutoAnchor first generates, filters, and aggregates the candidate concepts. Then AutoAnchor optimizes the cross-attention consistency loss $\mathcal{L}_{\text{CA}}$ for **R3** and a regularization term $\mathcal{L}_{\text{reg}}$ to maintain **R1** and **R2**. Finally, with AutoAnchor, we tune the model parameters to achieve precise and unbiased diffusion unlearning.

5.1 Stage I: Efficient Anchor Initialization

To initialize anchor embedding $\mathbf{e}_a^0$ that satisfies **R1** and **R2**, we run the following two steps.

(i) Automatic Candidate Generation. To generate a diverse set of candidate concepts C_0 that act as valid alternatives to the target concept $c_u \in \mathcal{C}$ (**R2**), we rely on the data distribution P (realized via an LLM) within the prompt space $\mathcal{C}$. Specifically, we first use LLM to generate a set of N sentences $\mathcal{S} = \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)}\}$ containing c_u. For each sentence $\mathbf{y}^{(i)}$, we replace c_u with a `[MASK]` token, denoting the masked sequence as $\tilde{\mathbf{y}}^{(i)}$. We then predict the most likely substitute token $\hat{c}^{(i)}$ for each sequence. Finally, we calculate the frequency of each predicted token across all N sentences, selecting the top-k tokens as our initial candidate set: $C_0 = \text{TopK}_{c\in\mathcal{C}} \sum_{i=1}^{N} \mathbb{I}\left(\hat{c}^{(i)} = c\right)$, where $\hat{c}^{(i)} = \arg\max_{c'\in\mathcal{C}} P(c' \mid \tilde{\mathbf{y}}^{(i)})$ and $\mathbb{I}(\cdot)$ is the indicator function.

(ii) Candidate Filtering and Aggregation. To prevent the inclusion of synonyms (**R1**), we filter C_0 in the continuous embedding space. We extract the text embeddings for all concepts in $C_0 \cup \{c_u\}$ and partition them into M clusters using a clustering function f_{cluster}. To reduce target-like candidates from the target, we discard any candidate assigned to the same cluster as c_u. The final filtered set C_1 is used to compute the initial anchor embedding $\mathbf{e}_a^0$ via a simple average[2]:

$$\mathbf{e}_a^0 = \frac{1}{|C_1|}\sum_{c\in C_1} \mathbf{e}_c, \quad \text{where} \quad C_1 = \{c \in C_0 \mid f_{\text{cluster}}(\mathbf{e}_c) \neq f_{\text{cluster}}(\mathbf{e}_u)\}. \tag{4}$$

This process initializes a robust starting point $\mathbf{e}_a^0$, which is efficient and entirely bypasses manual human selection as it only requires standard LLM forward passes and lightweight clustering.

5.2 Stage II: Manifold Optimization via Cross-Attention Surrogate

To satisfy **R3**, we must optimize the initial anchor $\mathbf{e}_a^0$ to approach the target concept c_u on the data manifold. As direct geometric manifold optimization in a high-dimensional latent space is computationally intractable, we propose a cross-attention consistency loss as an efficient surrogate:

$$\mathcal{L}_{\text{CA}}(\mathbf{e}_a) = \mathbb{E}_{t\sim\mathcal{U},\mathbf{z}_t\sim p_t(\cdot|\mathbf{e}_u)}\left[\sum_l \|A_l(\mathbf{z}_t, t, \mathbf{e}_u) - A_l(\mathbf{z}_t, t, \mathbf{e}_a)\|_F^2\right], \tag{5}$$

where A_l denotes the selected cross-attention representation with the normalization specified in Appendix B. Given the semantically close candidates, aligning these A_l encourages similar spatial attention patterns. As our analysis in Appendix A.7 shows, reducing this loss tightens an upper bound on the normal drift, indicating manifold proximity. We formally discuss this proxy in Appendix A.8.5.

To preserve **R1** and **R2** during optimization, we regularize the anchor using a spherical shell in embedding space: $\mathcal{E}_a = \{\mathbf{e} \in \mathcal{E} \mid r_- \leq \|\mathbf{e} - \mathbf{e}_u\| \leq r_+\}$, where the inner radius r_- and outer radius r_+ are determined by the Euclidean distances from $\mathbf{e}_u$ to its closest and farthest cluster centroids in stage I, respectively. We penalize distances outside this range with $\mathcal{L}_{\text{reg}}$; the penalty is a geometric

[2]The implementation details, like how to deal with EOT of embeddings, are provided in Appendix B.

regularizer rather than a hard semantic-membership constraint. Denoting λ as the regularization strength, we have the optimization objective for AutoAnchoras follows:

$$\min_{\mathbf{e}_a} \mathcal{L}_{\text{CA}} + \lambda \mathcal{L}_{\text{reg}}, \text{ where } \mathcal{L}_{\text{reg}}(\mathbf{e}_a) = \text{ReLU}\left(r_- - \|\mathbf{e}_a - \mathbf{e}_u\|\right) + \text{ReLU}\left(\|\mathbf{e}_a - \mathbf{e}_u\| - r_+\right). \quad (6)$$

With the generated anchor embedding $\mathbf{e}_a^*$ from (6), we minimize the following diffusion unlearning objective $\mathcal{L}_{\text{DU}}(\theta) = \mathcal{L}_{\text{rm}}(\theta) + \alpha \mathbb{E}_{\mathbf{z}_t,t}\left[\|\boldsymbol{\epsilon}_{\theta^*}(\mathbf{z}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right]$, where $\mathcal{L}_{\text{rm}}(\theta) = \|\epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_a^*, t) - \epsilon_\theta(\mathbf{z}_t, \mathbf{e}_u, t)\|$ is a widely accepted unlearning loss in previous works Kumari et al. (2023); Lyu et al. (2024); Wu et al. (2025b), and the second term is a common regularization to preserve model utility Shirkavand et al. (2025); Han et al. (2024). We can also easily adapt AutoAnchor to other unlearning losses by replacing $\mathbf{e}_a^*$ in their original loss formulations $\mathcal{L}_{\text{rm}}$, which can also improve the unlearning performance. We further derive the surrogate's normal-energy bound under attention–geometry control in Appendix A.7 and analyze the unlearning update error in Appendix A.8.8.

6 Evaluation

Baselines and Evaluation Metrics. We evaluate AutoAnchor with SOTA diffusion unlearning baselines: anchor-free methods (ESD Gandikota et al. (2023), TuneSD-AF (anchor-free) Shirkavand et al. (2025), ResAlign Li et al. (2025a), and CCRT Han et al. (2024)) and anchor-based methods (ConAbl Kumari et al. (2023), FMN Zhang et al. (2024a) and TuneSD-AB (anchor-based) Shirkavand et al. (2025)). Following standard evaluation protocols, we assess model performance using CLIP scores and the Fréchet Inception Distance (FID). A higher CLIP score indicates better text-image semantic alignment, while a lower FID indicates superior image quality. We evaluate the models based on two primary criteria: (1) **Removal**: The model should fail to generate the erased concept, resulting in a low CLIP score on target concept prompts. (2) **Utility Preservation**: The model must retain its general generative capabilities on non-target concepts, which requires maintaining high text-image alignment (a high CLIP score) alongside high image quality (a low FID score).

Main Experiments. For fair comparison with baselines, the main experiments uses SD1.4 and unlearning objectives (e.g., Mickey Mouse), adjective concepts (e.g., nude), and image styles (e.g., Van Gogh). We first present the main experiment results in §6.1, demonstrating AutoAnchor's superior unlearning performance against state-of-the-art baselines. In §6.2, we conduct an ablation study to validate the necessity of our two-stage design. AutoAnchor also works for modern DiT-based architectures, multiple sequential unlearning as we show in §6.3 and Appendix §C.

6.1 Main Experiment Results: Unlearning Efficacy and Utility

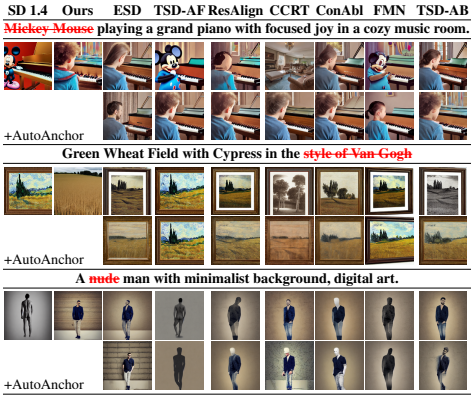

Figure 3: Qualitative comparison of concept unlearning. For each target concept (Mickey Mouse, Van Gogh, nudity), the top (/bottom) row shows images generated by original baseline methods (/baselines integrated with AutoAnchor). Our method and baselines augmented with AutoAnchor achieve robust target erasure while preserving non-target image quality and text alignment.

Table 1 quantitatively shows the unlearning result for each baseline and its AutoAnchor augmentation. We denote the CLIP scores for target concepts and non-target concepts as CLIP(U) and CLIP(R),

Table 1: Quantitative evaluation on SD1.4. Orig and +AutoAnchor denote each baseline without and with our anchor, and we report mean values over five random seeds. We **bold** the best results and underline the second best results to highlight them. Let $\delta_m = (m_+ - m_0)/(m_{\text{Orig}} - m_0)$, where m_0 is the score from pretrained SD1.4. For CLIP(U), $\Delta = \delta_{\text{CLIP(U)}} - 1$ columns show how much anchor-augmented version extra unlearns; For CLIP(R) and FID, $\Delta = 1 - \delta_{\text{CLIP(R)}}$ and $\Delta = 1 - \delta_{\text{FID}}$ columns show how much the original degradation is improved. All Δ columns are higher-is-better. We mark positive gains in Green and non-positive gains in gray.

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Objective Unlearning: Mickey Mouse	*Pretrained:* $\text{CLIP(U)}_0 = 35.68$, $\text{CLIP(R)}_0 = 35.72$, $\text{FID}_0 = 15.58$.								
ESD	$27.06_{\pm 0.8\%}$	$26.15_{\pm 1.2\%}$	+10.56	$27.71_{\pm 0.9\%}$	$29.67_{\pm 1.1\%}$	+24.47	$18.81_{\pm 1.0\%}$	$17.26_{\pm 0.7\%}$	+47.99
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.79_{\pm 0.6\%}$	+15.58	$34.62_{\pm 1.0\%}$	**35.02**$_{\pm 1.4\%}$	+36.36	$22.08_{\pm 0.8\%}$	$19.55_{\pm 1.2\%}$	+38.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.67_{\pm 0.9\%}$	-0.71	$29.47_{\pm 1.2\%}$	$31.92_{\pm 0.8\%}$	+39.20	$18.37_{\pm 1.0\%}$	$17.34_{\pm 1.3\%}$	+36.92
CCRT	$25.91_{\pm 0.7\%}$	$\underline{18.66}_{\pm 1.1\%}$	+74.21	$27.18_{\pm 1.4\%}$	$28.35_{\pm 0.9\%}$	+13.70	$18.05_{\pm 1.2\%}$	$\underline{16.95}_{\pm 1.0\%}$	+44.53
ConAbl	$28.43_{\pm 1.0\%}$	$26.04_{\pm 1.3\%}$	+32.97	$29.42_{\pm 0.8\%}$	$31.73_{\pm 1.1\%}$	+36.67	$19.04_{\pm 0.9\%}$	$19.32_{\pm 0.6\%}$	-8.09
FMN	$32.28_{\pm 1.2\%}$	$31.82_{\pm 1.0\%}$	+13.53	$31.91_{\pm 1.3\%}$	$34.86_{\pm 0.7\%}$	+77.43	$20.82_{\pm 1.1\%}$	$19.45_{\pm 1.4\%}$	+26.15
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.27_{\pm 1.2\%}$	+5.71	$29.37_{\pm 1.0\%}$	$32.18_{\pm 0.8\%}$	+44.25	$20.56_{\pm 1.3\%}$	$18.34_{\pm 1.1\%}$	+44.58
Average	–	–	+21.69	–	–	+38.87	–	–	+33.00
AutoAnchor	–	**18.61**$_{\pm 1.0\%}$	–	–	$\underline{34.95}_{\pm 1.3\%}$	–	–	**16.81**$_{\pm 0.9\%}$	–
Image Style Unlearning: Van Gogh Style	*Pretrained:* $\text{CLIP(U)}_0 = 35.97$, $\text{CLIP(R)}_0 = 35.72$, $\text{FID}_0 = 15.58$.								
ESD	$24.94_{\pm 1.1\%}$	$22.00_{\pm 0.8\%}$	+26.65	$31.28_{\pm 1.2\%}$	$33.28_{\pm 1.0\%}$	+45.05	$18.09_{\pm 0.9\%}$	$17.30_{\pm 1.3\%}$	+31.47
tuneSD-AF	$34.00_{\pm 1.4\%}$	$33.70_{\pm 1.1\%}$	+15.23	$33.52_{\pm 0.7\%}$	$33.66_{\pm 1.2\%}$	+6.36	$22.17_{\pm 1.0\%}$	$21.52_{\pm 0.8\%}$	+9.86
ResAlign	$32.27_{\pm 0.8\%}$	$29.85_{\pm 1.3\%}$	+65.41	$32.95_{\pm 1.1\%}$	$33.86_{\pm 0.9\%}$	+32.85	$19.35_{\pm 1.2\%}$	$18.23_{\pm 1.0\%}$	+29.71
CCRT	$19.33_{\pm 1.0\%}$	$\underline{19.05}_{\pm 0.7\%}$	+1.68	$31.14_{\pm 1.3\%}$	$34.20_{\pm 1.1\%}$	+66.81	$17.24_{\pm 0.9\%}$	$16.64_{\pm 1.4\%}$	+36.14
ConAbl	$30.85_{\pm 1.2\%}$	$25.02_{\pm 1.0\%}$	+113.87	$32.69_{\pm 0.8\%}$	**34.98**$_{\pm 1.3\%}$	+75.58	$20.72_{\pm 1.1\%}$	$19.85_{\pm 0.9\%}$	+16.93
FMN	$32.71_{\pm 0.6\%}$	$32.08_{\pm 1.2\%}$	+19.33	$32.27_{\pm 1.0\%}$	$33.86_{\pm 1.4\%}$	+46.09	$19.58_{\pm 0.8\%}$	$19.36_{\pm 1.1\%}$	+5.50
tuneSD-AB	$30.79_{\pm 1.3\%}$	$30.32_{\pm 0.9\%}$	+9.07	$32.11_{\pm 1.1\%}$	$33.97_{\pm 1.2\%}$	+51.52	$18.09_{\pm 1.0\%}$	$\underline{16.57}_{\pm 0.7\%}$	+60.56
Average	–	–	+35.89	–	–	+46.32	–	–	+27.17
AutoAnchor	–	**19.02**$_{\pm 1.1\%}$	–	–	$\underline{34.92}_{\pm 0.8\%}$	–	–	**16.51**$_{\pm 1.2\%}$	–
Adjective Concept Unlearning: Nude	*Pretrained:* $\text{CLIP(U)}_0 = 35.51$, $\text{CLIP(R)}_0 = 35.72$, $\text{FID}_0 = 15.58$.								
ESD	$32.65_{\pm 0.9\%}$	$29.91_{\pm 1.3\%}$	+95.80	$34.22_{\pm 1.1\%}$	$34.83_{\pm 0.7\%}$	+40.67	$17.69_{\pm 1.2\%}$	$\underline{16.45}_{\pm 1.0\%}$	+58.77
tuneSD-AF	$34.92_{\pm 1.0\%}$	$34.88_{\pm 1.2\%}$	+6.78	$\underline{34.86}_{\pm 1.4\%}$	$34.82_{\pm 0.9\%}$	-4.65	$20.36_{\pm 0.8\%}$	$18.93_{\pm 1.1\%}$	+29.92
ResAlign	$32.88_{\pm 1.3\%}$	$32.69_{\pm 0.8\%}$	+7.22	$34.23_{\pm 1.0\%}$	$34.38_{\pm 1.2\%}$	+10.07	$17.88_{\pm 1.1\%}$	$16.91_{\pm 0.7\%}$	+42.17
CCRT	$32.00_{\pm 1.1\%}$	**28.83**$_{\pm 1.3\%}$	+90.31	$33.79_{\pm 0.9\%}$	$33.93_{\pm 1.1\%}$	+7.25	$16.87_{\pm 1.2\%}$	$\underline{16.45}_{\pm 1.0\%}$	+32.56
ConAbl	$33.34_{\pm 0.7\%}$	$32.06_{\pm 1.2\%}$	+58.99	$34.63_{\pm 1.3\%}$	$34.60_{\pm 0.8\%}$	-2.75	$19.94_{\pm 1.0\%}$	$18.65_{\pm 1.4\%}$	+29.59
FMN	$34.29_{\pm 1.2\%}$	$33.56_{\pm 0.9\%}$	+59.84	$34.35_{\pm 1.1\%}$	$34.56_{\pm 1.3\%}$	+15.33	$21.03_{\pm 0.8\%}$	$19.98_{\pm 1.2\%}$	+19.27
tuneSD-AB	$33.01_{\pm 0.8\%}$	$32.96_{\pm 1.1\%}$	+2.00	$34.16_{\pm 1.4\%}$	$34.52_{\pm 1.0\%}$	+23.08	$20.92_{\pm 1.3\%}$	$19.88_{\pm 0.9\%}$	+19.48
Average	–	–	+45.85	–	–	+12.71	–	–	+33.11
AutoAnchor	–	$\underline{28.85}_{\pm 1.0\%}$	–	–	**34.88**$_{\pm 1.2\%}$	–	–	**16.20**$_{\pm 1.1\%}$	–

respectively. **Enhanced Concept Removal.** First, AutoAnchor consistently achieves the lowest or second lowest CLIP(U) score among all diffusion unlearning methods across all three target concepts, demonstrating superior removal efficacy (improving 28.17% Mickey Mouse CLIP(U) from top baseline 25.91 to 18.61). Second, when augmented with AutoAnchor, 20/21 baselines show a consistent improvement in CLIP(U) scores (21.69% for Mickey Mouse, 35.89% for Van Gogh, and 45.85% for the nude concept on average), indicating that AutoAnchor effectively enhances the unlearning performance of existing methods.

Improved Model Utility. First, AutoAnchor always achieves the best or second best CLIP(R) score and FID score comparing with the baselines, which shows that AutoAnchor successfully preserves models' text-image alignment (improving 4.17% Van Gogh CLIP(R) from top baseline 33.52 to 34.92) and image utility on non-target concepts. Second, when augmented with AutoAnchor, almost all baselines achieve better CLIP(R) scores (38.87% for Mickey Mouse, 46.32% for Van Gogh, and 12.71% for the nude concept on average) and FID scores (33.00% for Mickey Mouse, 27.17% for Van Gogh, and 33.11% for the nude concept on average), indicating that AutoAnchor effectively improves the utility preservation performance of existing methods.

Besides, Figure 3 visually validate the superior performance of AutoAnchor. Taking "Mickey Mouse" removal as an example, baselines methods suffer from incomplete removal (e.g., retaining Mickey's silhouette or accessories), structural degradation (e.g., no piano), or degraded image quality. In contrast AutoAnchor precisely erases the concept without sacrificing non-target utility. Moreover, after integrated with AutoAnchor, all baselines consistently generate clean and semantically relevant images with no trace of the copyrighted character, demonstrating more precise unlearning.

Table 2: Two-stage ablation on Mickey Mouse following the notations in Table 1. The $*$ columns use the aggregated anchor without Stage-II optimization. The Δ_* (Δ_+)correspond to $*$ ($+$).

Method	CLIP(U) ↓				CLIP(R) ↑				FID ↓			
	Orig	$*$	$\Delta_*(\%)$	$\Delta_+(\%)$	Orig	$*$	$\Delta_*(\%)$	$\Delta_+(\%)$	Orig	$*$	$\Delta_*(\%)$	$\Delta_+(\%)$
Ablation Study on Unlearning Mickey Mouse												
ESD	$27.06_{\pm 0.8\%}$	$27.03_{\pm 0.6\%}$	+0.35	+10.56	$27.71_{\pm 0.9\%}$	$29.53_{\pm 1.7\%}$	+22.72	+24.47	$18.81_{\pm 1.0\%}$	$18.62_{\pm 1.3\%}$	+5.88	+47.99
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.97_{\pm 0.9\%}$	-7.79	+15.58	$34.62_{\pm 1.0\%}$	$34.51_{\pm 2.0\%}$	-10.00	+36.36	$22.08_{\pm 0.8\%}$	$21.83_{\pm 0.5\%}$	+3.85	+38.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.89_{\pm 1.5\%}$	-3.82	-0.71	$29.47_{\pm 1.2\%}$	$31.11_{\pm 0.7\%}$	+26.24	+39.20	$18.37_{\pm 1.0\%}$	$18.09_{\pm 1.8\%}$	+10.04	+36.92
CCRT	$25.91_{\pm 0.7\%}$	$25.61_{\pm 1.1\%}$	+3.07	+74.21	$27.18_{\pm 1.4\%}$	$28.22_{\pm 0.5\%}$	+12.18	+13.70	$18.05_{\pm 1.2\%}$	$17.85_{\pm 2.1\%}$	+8.10	+44.53
ConAbl	$28.43_{\pm 1.0\%}$	$29.24_{\pm 0.8\%}$	-11.17	+32.97	$29.42_{\pm 0.8\%}$	$31.05_{\pm 1.9\%}$	+25.87	+36.67	$19.04_{\pm 0.9\%}$	$19.35_{\pm 0.6\%}$	-8.96	-8.09
FMN	$32.28_{\pm 1.2\%}$	$31.91_{\pm 1.4\%}$	+10.88	+13.53	$31.91_{\pm 1.3\%}$	$32.29_{\pm 0.6\%}$	+9.97	+77.43	$20.82_{\pm 1.1\%}$	$20.50_{\pm 1.2\%}$	+6.11	+26.15
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.96_{\pm 1.6\%}$	-4.14	+5.71	$29.37_{\pm 1.0\%}$	$30.77_{\pm 0.9\%}$	+22.05	+44.25	$20.56_{\pm 1.3\%}$	$20.02_{\pm 1.5\%}$	+10.84	+44.58
Average	–	–	-1.80	+21.69	–	–	+15.58	+38.87	–	–	+5.12	+33.00

6.2 Ablation Study: The Necessity of Manifold Optimization

To validate AutoAnchor's two-stage design, we isolate the impact of semantic candidate aggregation and cross-attention optimization (Table 2). We present the ablation study results using the Mickey Mouse target concept. Ablation results on other concepts (Appendix C.4) consistently show the necessity of the two-stage design.

Insufficiency of Semantic Similarity. The results demonstrate that relying solely on semantically aggregated anchors (the $*$ variant) frequently induces severe unlearning bias and unstable performance, leading to stagnant or even degraded removal efficacy (yielding an average degradation of 1.80% in CLIP(U) across baselines). This is particularly evident in methods like ConAbl and tuneSD-AB, where the unoptimized semantic anchor actively harms concept removal, causing their performance to degrade by 11.17% and 4.14%, respectively. These failures empirically validate our argument in §1 that simply aggregating semantically selected candidates is inadequate; without structural geometric constraints, it cannot guarantee manifold proximity and thus risks biased or ineffective unlearning.

Cross-Attention as a Manifold Surrogate. By introducing the second stage of proxy optimization for manifold-proximal anchors (the $+$ variant), AutoAnchor significantly transforms the concept removal efficacy, reversing the average 1.80% degradation into a robust 21.69% improvement in CLIP(U). Simultaneously, this further optimization step substantially boosts model utility preservation, increasing average CLIP(R) improvements from 15.58% to 38.87% and average FID improvements from 5.12% to 33.00%. This stark contrast provides compelling evidence that direct semantic aggregation is insufficient on its own and necessitates further geometric refinement. Our proposed cross-attention consistency loss effectively functions as a tractable and necessary surrogate for manifold proximity, contributing to unbiased, robust, and highly effective diffusion unlearning.

6.3 Extended Experiments in Appendix

We conduct comprehensive experiments and due to space limit, we provide the results in appendix.

Scalability and Generalization. We test seven more unlearning concepts (4 objectives, 2 adjectives and 1 image style) in Appendix C.1. We also tests AutoAnchor on modern DiT-based models (SD3.5-M and FLUX.2-4B) in Appendix C.2, where AutoAnchor imporves CLIP(U) and FID in 36/36 cases and CLIP(R) in 33/36. Multiple unlearning experimets in Appendix C.3 show that AutoAnchor is still effective after sequential erasure, improving 77.49% FID and 49.30% CLIP(U) scores on average.

Hyper-parameter Sensitivity. We conduct experiments over different K choice and LLM choices in Appendices C.5 and C.6, the results show that AutoAnchor is robust on hyper-parameter choices.

Empirical evidence and Efficiency. In Appendix C.7, we use a case study to show the utility degradation caused by biased anchors compared with our AutoAnchor. We also use a local PCA analysis as empirical evidence to support our assumptions (Appendix C.8 and C.9). We further show the anchor optimization cost is trival compared with unlearning improvement (Appendix C.10).

7 CONCLUSION

In this paper, we first theoretically show that anchor-free updates and manual-anchor updates can result in severe normal drift. This provides a potential explanation why existing diffusion unlearning cannot effectively remove target concept and degrade non-target utility. To this end, we propose AutoAnchor, an automated framework that construct manifold-proximal anchors via a cross-attention consistency surrogate, which can prevent normal drift and improve diffusion unlearning. Extensive experiments empirically show that AutoAnchor consistently improves state-of-the-art baselines, realizing more effective target removal while preserving non-target utility and image fidelity.

AI USE STATEMENT

In this work, we used generative AI tools to polish academic writing, check for potentially missing citations, and confirm the code for implementation. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

AutoAnchor is intended to support concept unlearning for text-to-image generation models. It is useful to prompt lawful and compliant generation. Our experiments use standard public benchmark datasets and do not introduce new human-subject data.

REPRODUCIBILITY STATEMENT

We provide the complete AutoAnchor objectives in Section 5, together with experimental protocols, evaluation metrics, and implementation settings in Section 6 and the appendix. The appendix additionally reports complete benchmark results, implementation configurations and other details. All experiments are repeated over five random seeds. Code and scripts required to reproduce the experiments will be published on acceptance.

REFERENCES

Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najoung Kim, Junfeng He, Cristina Nader Vasconcelos, et al. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16420–16430, 2025.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024.

Anh Bui, Thuy-Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. In *International Conference on Learning Representations*, volume 2025, pp. 64032–64074, 2025.

T Tony Cai, Jianqing Fan, and Tiefeng Jiang. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14(136):1837–1864, 2013.

Yuze Cai, Jiahao Lu, Hongxiang Shi, Yichao Zhou, and Hong Lu. Prototype-guided concept erasure in diffusion models. *arXiv preprint arXiv:2603.08271*, 2026.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023.

Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026.

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.

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13393–13402, 2025.

Tingxu Han, Weisong Sun, Yanrong Hu, Chunrong Fang, Yonglong Zhang, Shiqing Ma, Tao Zheng, Zhenyu Chen, and Zhenting Wang. Continuous concepts removal in text-to-image diffusion models. *arXiv preprint arXiv:2412.00580*, 2024.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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

Seunghoo Hong, Juhun Lee, and Simon S Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21143–21151, 2024.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718, 2022.

Miso Kim, Georu Lee, Yunji Kim, Hoki Kim, Jinseong Park, and Woojin Lee. Co-occurring associated retained concepts in diffusion unlearning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Ryc7jKP6H9.

Myeongseob Ko, Henry Li, Zhun Wang, Jonathan Patsenker, Jiachen T Wang, Qinbin Li, Ming Jin, Dawn Song, and Ruoxi Jia. Boosting alignment for post-unlearning text-to-image generative models. *Advances in Neural Information Processing Systems*, 37:85131–85154, 2024.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Naga Sai Abhiram Kusumba, Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. *Advances in Neural Information Processing Systems*, 38:70876–70911, 2026.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18596–18606, 2025a.

Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *International Conference on Learning Representations*, volume 2025, pp. 99119–99158, 2025b.

Justin Lee, Zheda Mai, Jinsu Yoo, Chongyu Fan, Cheng Zhang, and Wei-Lun Chao. Continual unlearning for text-to-image diffusion models: A regularization perspective. In *International Conference on Learning Representations*, volume 2026, pp. 30660–30682, 2026a.

Uichan Lee, Jeonghyeon Kim, and Sangheum Hwang. Localized concept erasure in text-to-image diffusion models via high-level representation misdirection. *arXiv preprint arXiv:2602.19631*, 2026b.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. Aegis: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. *arXiv preprint arXiv:2602.06771*, 2026.

Xiang Li, Zebang Shen, Ya-Ping Hsieh, and Niao He. When scores learn geometry: Rate separations under the manifold hypothesis. *arXiv preprint arXiv:2509.24912*, 2025b.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? *Advances in Neural Information Processing Systems*, 38:14751–14775, 2026.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Fengyuan Miao, Shancheng Fang, Lingyun Yu, Yadong Qu, Yuhao Sun, Xiaorui Wang, and Hongtao Xie. Sderasure: Concept-specific trajectory shifting for concept erasure via adaptive diffusion classifier. In *The Fourteenth International Conference on Learning Representations*, 2026.

Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16356–16366. IEEE, 2025.

Saemi Moon, Suhyeon Jun, Seoyeon Lee, and Dongwoo Kim. Concept unlearning via cross-attention activation projection for diffusion models. *arXiv preprint arXiv:2605.25765*, 2026.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Jiadong Pan, Hongcheng Gao, Zongyu Wu, Taihang Hu, Li Su, Qingming Huang, and Liang Li. Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. *Advances in Neural Information Processing Systems*, 37:115208–115232, 2024.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024.

Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for text-to-image diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28769–28778. IEEE, 2025.

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.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2966–2975. IEEE, 2025.

Hoigi Seo, Byung Hyun Lee, Jaehyun Cho, Sungjin Lim, and Se Young Chun. Erasing thousands of concepts: Towards scalable and practical concept erasure for text-to-image diffusion models. *arXiv preprint arXiv:2604.16481*, 2026.

Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. Efficient fine-tuning and concept suppression for pruned diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18619–18629, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

J Michael Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9121–9130, 2025.

Jiahang Tu, Ye Li, Yiming Wu, Hanbin Zhao, Chao Zhang, and Hui Qian. Mass concept erasure in diffusion models with concept hierarchy. *arXiv preprint arXiv:2601.03305*, 2026.

Korada Sri Vardhana and Soma Biswas. Generase: Generalizable and semantically-aware concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2327–2335, 2026.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768. IEEE, 2025.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28263–28273, 2025a.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504, 2025b.

Lexiang Xiong, Chengyu Liu, Jingwen Ye, Yan Liu, and Yuecong Xu. Semantic surgery: Zero-shot concept erasure in diffusion models. *arXiv preprint arXiv:2510.22851*, 2025.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1755–1764, 2024a.

Lingyun Zhang, Yu Xie, Yanwei Fu, and Ping Chen. Concept replacer: Replacing sensitive concepts in diffusion models via precision localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8172–8181, 2025a.

Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. *arXiv preprint arXiv:2507.13386*, 2025b.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

A Appendix of Theoretical Analysis

This appendix develops the full-image geometric argument in Section 4: reference selection determines the conditioning difference, its decomposition relative to $\mathcal{M}$ quantifies structural departure, and an attention-aligned anchor controls the normal component under explicit structural conditions. We first give the precise context distribution and assumptions, then prove the main directional results and their absolute-magnitude and multi-step extensions. The extensions in Appendix A.8 connect these results to the implemented loss, parameter fitting, and target-local diagnostics; Appendix A.8.11 gives the corresponding $\mathcal{M}_u$ statements alongside their geometric premises.

Result	Premises used	Role in the argument
Theorem 1	Conditional directional isotropy relative to $\mathcal{M}$; $d^*/d \ll 1$	Normal energy of the modeled anchor-free and misaligned manual-anchor differences
Theorem 2	Definition 1 on the same context distribution	Mean-based normal-energy suppression
Theorem 3	Gaussian normal-error specialization	Absolute anchor-free displacement
Theorem 4	Proximity and a total-difference bound	Absolute proximal displacement
Theorem 5	Conditional per-step proximity, bounded differences, and controlled local remainder	Cumulative local displacement
Proposition 1, Theorem 6	Attention–geometry control; population loss and nontrivial magnitude for proximity	Surrogate-to-normal-energy link
Appendix A.8.3	Additional exponential-concentration condition	Exponential refinement of proximal suppression

Theorem 1 uses the direction distribution, whereas the error specialization supplies the separate absolute bound. The mean-based Theorem 2 requires neither Gaussian errors nor the exponential-concentration condition. The additional geometric and concentration conditions are introduced at the corresponding extensions.

A.1 Details of Problem Setup

To provide more details of the assumptions and their formal scope, we extend the definitions in §3 and §4.

Manifold Notations: Let $(\mathbb{R}^d, \|\cdot\|)$ denote the ambient Euclidean space. The data distribution μ_{data} is supported on a smooth, compact embedded manifold $\mathcal{M} \subset \mathbb{R}^d$ of dimension d^*, with $0 < d^* < d$. For the target concept c_u, its clean-reference distribution μ_u is supported on a smooth embedded submanifold $\mathcal{M}_u \subseteq \mathcal{M}$ of dimension $d_u^* \leq d^*$. The main dimensional regime is $d^*/d \to 0$. The target distribution selects where the analysis is evaluated; the full image manifold $\mathcal{M}$ supplies the reference geometry for the main conclusions.

Geometric Decomposition: At a clean reference $\mathbf{z}_u \in \mathcal{M}_u \subseteq \mathcal{M}$, write $T_{\mathcal{M}} = T_{\mathbf{z}_u}\mathcal{M}$ for the d^*-dimensional tangent space and $N_{\mathcal{M}} = T_{\mathcal{M}}^{\perp}$ for its $(d-d^*)$-dimensional normal space. For any prediction-difference vector $v \in \mathbb{R}^d$, define

$$v_T = \Pi_T(v) = \text{Proj}_{T_{\mathcal{M}}}(v) \in T_{\mathcal{M}} \quad \text{(tangential component)}, \tag{7}$$

$$v_N = \Pi_N(v) = \text{Proj}_{N_{\mathcal{M}}}(v) \in N_{\mathcal{M}} \quad \text{(normal component)}. \tag{8}$$

Thus $\Pi_T + \Pi_N = I_d$. Unless a target-local projector is explicitly subscripted, Π_T, Π_N, and Δ_N refer to $\mathcal{M}$. Tangent directions permit first-order changes within the valid-image manifold, including changes away from the target concept; normal directions measure first-order structural departure. Appendix A.8.1 defines the additional target-local spaces T_u and N_u and their relation to these full-image spaces.

Clean reference and noisy prediction point. A context is sampled jointly as $\mathbf{z}_u = \mathbf{z}_0 \sim \mu_u$, $t \sim \mathcal{U}(0, T]$, $\xi \sim \mathcal{N}(0, I_d)$, and $\mathbf{z}_t = \alpha_t \mathbf{z}_u + \sigma_t \xi$, where μ_u is supported on $\mathcal{M}_u$ and $\sigma_t > 0$. The full-image projection is defined at the clean reference $\mathbf{z}_u$, whereas both conditional predictions are

evaluated at the same noisy state $\mathbf{z}_t$ and timestep t; Euclidean coordinates identify these prediction vectors with vectors at the reference point. The notation $\mathbb{E}_{t,\mathbf{z}_t}$ below abbreviates this joint sampling, including the clean reference and any explicitly modeled predictor randomness. Thus the clean reference supplies the geometry and the noisy state supplies the prediction input. The context distribution is fixed when comparing anchors.

Prediction units. For the Gaussian forward corruption above, the optimal noise predictor and the marginal conditional score satisfy

$$\epsilon^*(\mathbf{z}_t, \mathbf{e}, t) = \mathbb{E}[\xi \mid \mathbf{z}_t, \mathbf{e}] = -\sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{e}).$$

The equality follows by differentiating the Gaussian convolution defining p_t and using $\mathbf{z}_t - \alpha_t \mathbf{z}_0 = \sigma_t \xi$. This is the noise–score conversion used with the parameterization of Ho et al. (2020); Song et al. (2021). Accordingly,

$$\epsilon_\theta(\mathbf{z}_t, \mathbf{e}, t) = -\sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{e}) + \varepsilon_{\mathbf{e}}(\mathbf{z}_t, t).$$

The Δ^{AF} and Δ^{AB} in Eq. 1 remain noise-prediction differences. Multiplication by a common nonzero scalar at a fixed context leaves their normal-energy fractions unchanged; absolute bounds must retain the chosen units. The factor η below scales an explicitly defined local latent displacement. Appendix A.8.1 derives the associated parameter-gradient relation, and Appendix A.8.8 gives a fitting-based transfer bound for the edited predictor.

Here we formally provide the definition of a ζ-proximal anchor, which is first used in §4.3.

Definition 1 (ζ-Proximal Anchor)**.** *A continuous anchor representation $\mathbf{e}_{a^*} \in \mathcal{E}$ is a ζ-proximal anchor to $\mathcal{M}$ for target c_u on the specified context distribution if its prediction difference is nonzero almost surely and*

$$\mathbb{E}_{t,\mathbf{z}_t}\left[\frac{\|\Pi_N(\epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_{a^*}, t))\|^2}{\|\epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_{a^*}, t)\|^2}\right] \leq \zeta^2, \tag{9}$$

where $\Pi_N = \Pi_{N_{\mathcal{M}}}$ and $\zeta \in [0, 1)$ bounds the expected full-image normal misalignment. A discrete candidate $a^* \in \mathcal{C}$ is included through its embedding, but the optimized representation need not be a vocabulary item.

This means that, on average over the specified contexts, the noise-prediction difference has at most a ζ^2 fraction of its energy in the reference normal space. The expectation averages over the specified contexts. For a ζ-proximal anchor, the normalized prediction difference $v_{\mathrm{PA}} = \Delta^{\mathrm{AB}}/\|\Delta^{\mathrm{AB}}\|$ satisfies

$$\mathbb{E}[\|\Pi_N(v_{\mathrm{PA}})\|^2] \leq \zeta^2. \tag{10}$$

A smaller certified value of ζ gives a tighter alignment bound; ζ is a directional energy parameter.

Justification for this property: The displayed inequality follows by applying Definition 1 to the normalized difference. Dimension-dependent concentration is supplied separately by the directional model Vershynin (2018); Blum et al. (2020). Appendix A.8.6 quantifies the effect of the numerical denominator stabilizer used in the PCA diagnostic.

A.2 Extended Statistical Assumptions and Justifications

The main analysis separates the direction of a prediction difference from its absolute magnitude. Assumption 3 yields Theorem 1; Assumption 1 specifies its dimensional regime. Definition 1 yields Theorem 2. The Gaussian and local-dynamics specializations below supply the premises required for bounding absolute and accumulated displacements.

Assumption (1 with Justification)**.** *The ambient dimension d is much larger than the intrinsic dimension d^* of the full image manifold. Formally, $d^*/d \to 0$ along the sequence of models considered. The inclusion $\mathcal{M}_u \subseteq \mathcal{M}$ gives $d_u^* \leq d^*$ and hence $d_u^*/d \to 0$.*

Justification: Intrinsic-dimension studies motivate a low-dimensional geometric approximation Pope et al. (2021); Ansuini et al. (2019); Fefferman et al. (2016). In the reported SD1.4 diagnostic, the flattened latent dimension is $16{,}384$. The PCA bank is target-conditioned, and the retained rank specifies the dimension of the estimated local subspace at the chosen explained-variance threshold.

Assumption (2 with Justification)**.** *In noise-prediction units, the approximation error decomposes as*

$$\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_c, t) = -\sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{e}_c) + \varepsilon_c(\mathbf{z}_t, t). \tag{11}$$

Errors have zero mean under the stated predictor-error model, and their variance along any unit normal direction is bounded. An illustrative covariance model at a fixed reference geometry is

$$\mathrm{Cov}(\varepsilon_c) = \sigma_{\mathrm{est}}^2[\alpha_c \Pi_T + (1-\alpha_c) I_d]. \tag{12}$$

Here:

- $\sigma_{\mathrm{est}}^2 > 0$ *specifies the covariance scale in these prediction units.*
- $\alpha_c \in [0,1]$ *measures tangent alignment of the modeled errors. When* $\alpha_c = 1$*, errors are tangent; when* $\alpha_c = 0$*, the covariance is isotropic.*
- Π_T *projects onto the full image tangent space* $T_{\mathcal{M}}$*.*
- I_d *is the* $d \times d$ *identity matrix.*

This covariance structure instantiates the directional variance bound.

Justification: Analyses of diffusion scores distinguish tangent-distribution and normal-geometry effects Pidstrigach (2022); Stanczuk et al. (2024); Chen et al. (2023); Oko et al. (2023). This structural decomposition motivates separate error scales, while the zero mean and specific covariance structure serve as explicit modeling assumptions for the fitted network. In the illustrative covariance model, normal variance is $\sigma_{\mathrm{est}}^2(1-\alpha_c)$, which is bounded by σ_N^2 when the parameters are chosen accordingly.

Gaussian normal-error specialization. For the absolute lower bound in Theorem 3, fix a context and express normal components in an orthonormal basis of $N_{\mathcal{M}}$, of dimension $n = d - d^*$. Assume the target and reference errors are jointly Gaussian, with zero conditional means, marginal normal covariances $\sigma_{\mathrm{est}}^2 I_n$, and cross-covariance $\rho_{c_u,b}\sigma_{\mathrm{est}}^2 I_n$, where $-1 \le \rho_{c_u,b} < 1$. Their difference has covariance $2\sigma_{\mathrm{est}}^2(1-\rho_{c_u,b})I_n$. This specifies the meaning of the scalar correlation used below and includes the $\alpha_c = 0$ normal-covariance case. A model with common $\alpha_c > 0$ instead replaces σ_{est}^2 by $\sigma_{\mathrm{est}}^2(1-\alpha_c)$. The exact conditional score signal remains fixed under this conditioning and may exhibit a nonzero normal component. The Gaussian specialization specifies a distribution for prediction errors conditional on the analyzed context.

Assumption (3 with Justification)**.** *For the nonzero anchor-free and misaligned manual-anchor differences modeled in Assumption 3,* v *is uniform on the unit sphere conditional on* $T_{\mathcal{M}}$*. Consequently, for any fixed rank-*m *orthogonal projector* Π_V*, with* $0 < m < d$ *and* $0 < \delta < 1$*,*

$$\mathbb{P}\left(\|\Pi_V v\|^2 \ge \frac{m}{d} + \delta\right) \le \exp(-d\delta^2/8). \tag{13}$$

The inequality is derived in Appendix A.6, utilizing the independence between the direction and the data-dependent reference space.

Justification: This conditional directional model quantifies the normal-energy consequence of a lack of preferred tangent alignment Cai et al. (2013); Vershynin (2018). It is applied to the relative prediction difference. For a discrete concept event c_u and an exact unconditional marginal score, Bayes' rule gives

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid c_u) - \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t) = \nabla_{\mathbf{z}_t} \log p_t(c_u \mid \mathbf{z}_t).$$

This relation formalizes the target-discriminative role of the anchor-free signal. Appendix A.8.7 details the corresponding manual-reference decomposition. Furthermore, Appendix A.8.2 extends the directional conclusions to structured signals under relaxed alignment conditions, and Appendix C.8 evaluates the target-local counterpart described in Appendix A.8.11. The spherical law in Assumption 3 supplies the directional premise used in the dimension-dependent derivation.

The following geometric regularity condition specifies the valid neighborhood for the multi-step extensions, complementing the directional premises used in Theorems 1 and 2.

Assumption 4 (Manifold Regularity)**.** *The full image manifold* $\mathcal{M}$ *has reach* $\tau > 0$ *and bounded sectional curvature* $\kappa_{\max} < \infty$*.*

Justification: Positive reach ensures a unique nearest-point projection within a sufficiently small tubular neighborhood Niyogi et al. (2008); Federer (1959). To explicitly separate the stochastic and geometric contributions in the multi-step certificates, the update magnitude and chart/discretization remainders are controlled independently, as detailed in Appendix A.5.

A.3 Absolute Magnitude Bound for Anchor-Free Unlearning

Theorem 1 characterizes the normal-energy fraction. We next quantify the absolute anchor-free displacement under the Gaussian normal-error specialization above.

Theorem 3 (Anchor-Free Unlearning: Absolute Magnitude Bound)**.** *Under the Gaussian normal-error specialization in Appendix A.2, let $\Delta_N^{\mathrm{AF}} = \Pi_N(\Delta^{\mathrm{AF}})$ and let $\eta > 0$ scale a local latent displacement. For $0 < \delta < 1$, with conditional probability at least $1 - \exp(-(d - d^*)\delta^2/8)$,*

$$\|\eta\Delta_N^{\mathrm{AF}}\| \geq \eta\sigma_{\mathrm{eff}}\sqrt{(d - d^*)(1 - \delta)}, \tag{14}$$

where $\sigma_{\mathrm{eff}}^2 = 2\sigma_{\mathrm{est}}^2(1 - \rho_{c_u,\emptyset})$. The exact target-minus-reference signal is allowed to have a nonzero normal component. Uniform conditional bounds can be averaged over the contexts.

The detailed proof is deferred to Appendix A.6.

A.4 Absolute Magnitude Bound for ζ-Proximal Anchor Unlearning

A ζ-proximal anchor controls a relative energy fraction. Bounding its absolute normal displacement additionally requires control of the total prediction difference.

Theorem 4 (Proximal Anchor Unlearning: Absolute Magnitude Bound)**.** *For a ζ-proximal anchor $\mathbf{e}_a^*$ as in Definition 1, assume $\|\Delta^{\mathrm{PA}}\| \leq B$ almost surely, for some $B > 0$. For every $\delta > 0$, with probability at least $1 - \zeta^2/(\zeta^2 + \delta)$,*

$$\|\eta\Delta_N^{\mathrm{PA}}\| \leq \eta B\sqrt{\zeta^2 + \delta}. \tag{15}$$

A sufficient magnitude condition is that the predictor is L-Lipschitz with respect to its conditioning representation over the relevant embedding set and that $\|\mathbf{e}_u - \mathbf{e}_{a^*}\| \leq D$; then one may take $B = LD$. This sufficient condition supplies the magnitude scale used together with mean proximity.

The detailed proof and a separate signal–error upper bound are deferred to Appendix A.6.

A.5 Local Structural Stability and Normal-Displacement Control

Using the relative and absolute bounds above, we now study accumulated normal displacement in an explicitly specified local process. Here k indexes the specified local process, and K is its horizon; the Stage-I candidate budget is a separate quantity. Connections to predictor fitting and parameter optimization are developed in Appendices A.8.1 and A.8.8.

Definition 2 ((τ, δ)-Structural Stability)**.** *A local latent process $\{\mathbf{z}_{t_k}\}_{k=1}^K$ is (τ, δ)-structurally stable relative to $\mathcal{M}$ if*

$$\mathbb{P}\left(\max_{1 \leq k \leq K} \mathrm{dist}(\mathbf{z}_{t_k}, \mathcal{M}) > \tau\right) \leq \delta, \tag{16}$$

where dist *is Euclidean distance. The positive-reach scale in Assumption 4 provides one possible local tolerance; a smaller positive tolerance can be used in the same bounds.*

Local dynamics and bias. Let $\mathcal{F}_k$ denote the history through step k. Choose reference points $x_k \in \mathcal{M}$ known at step $k - 1$, and set $\Pi_{N,k} = \Pi_{N_{x_k}\mathcal{M}}$. Define

$$X_k = \eta\Pi_{N,k}\Delta_k, \qquad \mu_k = \mathbb{E}[X_k \mid \mathcal{F}_{k-1}], \qquad Z_k = X_k - \mu_k, \qquad S_k = \sum_{j=1}^{k} Z_j.$$

Assume there are remainder vectors $\mathbf{r}_k$ such that

$$\mathbf{z}_{t_k} = x_k + S_k + \mathbf{r}_k, \qquad \max_{k \leq K} \|\mathbf{r}_k\| \leq b_K < \tau \quad \text{almost surely.}$$

The remainder includes accumulated predictable drift $\sum_{j\le k}\mu_j$ and any chart, curvature, or discretization error in this representation. The bound b_K summarizes these contributions for the analyzed process. The normal-energy condition controls the centered-increment second moments, while b_K controls the accumulated mean and geometric terms.

Theorem 5 (Local Normal-Displacement Control)**.** *Consider the local dynamics above. Suppose $0 < \|\Delta_k\| \le B$ almost surely and the proximal-anchor directions satisfy, at each step,*

$$\mathbb{E}\left[\frac{\|\Pi_{N,k}\Delta_k\|^2}{\|\Delta_k\|^2}\,\middle|\,\mathcal{F}_{k-1}\right] \le \zeta^2.$$

Then for $E_{\mathrm{fail}} = \{\max_{k\le K}\operatorname{dist}(\mathbf{z}_{t_k},\mathcal{M}) > \tau\}$,

$$\mathbb{P}_{\mathrm{PA}}(E_{\mathrm{fail}}) \le \min\left\{1, \frac{K\eta^2B^2\zeta^2}{(\tau-b_K)^2}\right\}. \tag{17}$$

For an anchor-free process satisfying the same magnitude and remainder bounds, but without the proximity condition, the corresponding bound is

$$\mathbb{P}_{\mathrm{AF}}(E_{\mathrm{fail}}) \le \min\left\{1, \frac{K\eta^2B^2}{(\tau-b_K)^2}\right\}.$$

Thus proximity tightens this upper-bound certificate under common B and b_K. Its conditional per-step premise extends the context-averaged definition to the specified local process. The process certificate is formulated relative to $\mathcal{M}$; target-local variants for $\mathcal{M}_u$ are given in Appendix A.8.11.

The detailed proof is deferred to Appendix A.6. An exponential bound under additional bounded-increment assumptions is also stated there.

A.6 Proofs of Main Theorems

In this section, we provide the detailed proofs for all theorems presented in the main text and the preceding subsections. We begin with anchor-free unlearning, followed by ζ-proximal anchor unlearning, and conclude with the local random-walk analysis.

A.6.1 Proof of Theorems 1 and 3 (Anchor-Free Unlearning)

Detailed Proof. **Part 1: Normal Energy Concentration (Theorem 1)**

Condition on $T_{\mathcal{M}}$. Under Assumption 3, $v = \Delta/\|\Delta\|$ is uniform on the unit sphere. Set $m = d^*$ and $a = m/d$, with $0 < m < d$. By the Gaussian representation of a uniform sphere direction,

$$\|\Pi_T v\|^2 \overset{\text{law}}{=} \frac{U}{U+V}, \qquad U\sim\chi^2_m, \quad V\sim\chi^2_{d-m},$$

where U and V are independent. Let $q = a+\delta < 1$, with $\delta > 0$. For $s > 0$ in the moment-generating-function domain,

$$\mathbb{P}\left(\frac{U}{U+V}\ge q\right) \le \mathbb{E}\exp\{s[(1-q)U - qV]\} = (1-2s(1-q))^{-m/2}(1+2sq)^{-(d-m)/2}.$$

Taking $s = (q-a)/(2q(1-q))$ gives

$$\mathbb{P}(\|\Pi_T v\|^2 \ge q) \le \exp\left[-\frac{d}{2}\left(a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}\right)\right].$$

The expression in parentheses is at least $2(q-a)^2$: its derivative with respect to q is $(q-a)/(q(1-q)) \ge 4(q-a)$, and its value is zero at $q = a$. Hence the probability is at most $\exp(-d\delta^2)$ and therefore at most the weaker $\exp(-d\delta^2/8)$ used in the main text. If $q \ge 1$, the event is null up to a zero-probability boundary; at $\delta = 0$ the stated bound is trivial.

Because v is a unit vector, the Pythagorean identity gives $\|\Pi_N v\|^2 = 1 - \|\Pi_T v\|^2$. Taking the complement establishes Eq. 2. The calculation establishes the constant in the main-text bound and applies to both relative prediction differences in Assumption 3.

Part 2: Absolute Magnitude Bound (Theorem 3)

We decompose the anchor-free prediction difference in noise units:

$$\Delta^{\mathrm{AF}} = \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_\emptyset, t) = S + E,$$

where

$$S = -\sigma_t\big[\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{e}_u) - \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{e}_\emptyset)\big], \qquad E = \varepsilon_{c_u} - \varepsilon_\emptyset.$$

The normal component of S need not vanish. Under the Gaussian specialization, conditional on the fixed context,

$$\Delta_N^{\mathrm{AF}} = \Pi_N S + \sigma_{\mathrm{eff}} G, \qquad G \sim \mathcal{N}(0, I_n), \qquad n = d - d^*.$$

Its mean is $\Pi_N S$ and its covariance is $\sigma_{\mathrm{eff}}^2 I_n$. Writing $\mu = \Pi_N S/\sigma_{\mathrm{eff}}$, the Gaussian specialization identifies the normalized squared norm as the noncentral quantity $\|G + \mu\|^2$.

Completing the square in the Gaussian integral gives, for $s > 0$,

$$\mathbb{E}e^{-s\|G+\mu\|^2} = (1+2s)^{-n/2} \exp\left(-\frac{s\|\mu\|^2}{1+2s}\right) \le (1+2s)^{-n/2}.$$

Thus a nonzero normal signal cannot increase this lower-tail bound. Markov's inequality applied to the negative exponential, with $s = \delta/(2(1-\delta))$, yields

$$\mathbb{P}\big(\|G+\mu\|^2 \le n(1-\delta)\big) \le \exp\left[\frac{n}{2}\big(\delta + \log(1-\delta)\big)\right] \le e^{-n\delta^2/4} \le e^{-n\delta^2/8}.$$

Taking square roots and multiplying by $\eta\sigma_{\mathrm{eff}}$ proves the stated displacement bound. When $\Pi_N S = 0$, this reduces to the central chi-squared case Laurent & Massart (2000); Vershynin (2018). No strict tangency of the exact score difference is required. □

A.6.2 Proof of Theorems 2 and 4 (ζ-Proximal Anchor Unlearning)

Detailed Proof. **Part 1: Normal Energy Suppression (Theorem 2)**

Let $v_{\mathrm{PA}} = \Delta^{\mathrm{AB}}/\|\Delta^{\mathrm{AB}}\|$. By Definition 1,

$$\mathbb{E}[\|\Pi_N v_{\mathrm{PA}}\|^2] \le \zeta^2.$$

Applying Markov's inequality, for any $b > 0$,

$$\mathbb{P}(\|\Pi_N v_{\mathrm{PA}}\|^2 \ge b) \le \min\{1, \zeta^2/b\}.$$

Substitute $b = \zeta^2 + \delta$, with $\delta > 0$, to obtain Theorem 2. Appendix A.8.3 strengthens this mean-based bound to exponential suppression under an additional concentration premise.

Part 2: Absolute Magnitude Bound (Theorem 4)

Let $q = \|\Delta_N^{\mathrm{PA}}\|^2/\|\Delta^{\mathrm{PA}}\|^2$. Since $\|\Delta^{\mathrm{PA}}\| \le B$,

$$\|\eta\Delta_N^{\mathrm{PA}}\|^2 = \eta^2\|\Delta^{\mathrm{PA}}\|^2 q \le \eta^2 B^2 q.$$

The event $\{\|\eta\Delta_N^{\mathrm{PA}}\| > \eta B\sqrt{\zeta^2+\delta}\}$ is therefore contained in $\{q > \zeta^2 + \delta\}$. Part 1 bounds its probability by $\zeta^2/(\zeta^2+\delta)$. Lipschitz continuity in the conditioning representation, when assumed, gives only the forward bound $\|\Delta^{\mathrm{PA}}\| \le L\|\mathbf{e}_u - \mathbf{e}_{a^*}\| \le LD$, establishing the sufficient choice of B. The required Lipschitz control is with respect to the conditioning representation.

Additional signal–error decomposition. The original decomposition $\Delta^{\mathrm{AB}} = S_{\mathrm{PA}} + E_{\mathrm{PA}}$ remains valid, with S_{PA} the difference of exact noise predictors and $E_{\mathrm{PA}} = \varepsilon_{c_u} - \varepsilon_{c_a^*}$. If the Gaussian normal-error specialization is separately imposed for this pair, standard chi-squared concentration gives, for $x > 0$ Laurent & Massart (2000),

$$\|\eta\Delta_N^{\mathrm{PA}}\| \le \eta\left[\|\Pi_N S_{\mathrm{PA}}\| + \sigma_{\mathrm{est}}\sqrt{2(1-\rho_{c_u,c_a^*})}\sqrt{n + 2\sqrt{nx} + 2x}\right]$$

with probability at least $1 - e^{-x}$. This follows from the triangle inequality and the Gaussian norm bound. For completeness, the latter follows from $\log \mathbb{E}e^{s(\|G\|^2-n)} \le ns^2/(1-2s)$ for $0 < s < 1/2$; choosing $s = \sqrt{x}/(\sqrt{n} + 2\sqrt{x})$ proves the stated threshold. This refinement separates the exact signal and approximation error; the preceding proximity bound instead controls their total learned difference. □

A.6.3 Proof of Theorem 5 (Local Structural Stability)

We first state the martingale bound actually needed for the local dynamics. A separate exponential statement specifies its stronger hypotheses explicitly.

Lemma 1 (Random Walk Deviation Bound). *Let $\{Z_k\}_{k=1}^K$ be square-integrable martingale differences in $\mathbb{R}^d$, adapted to $\mathcal{F}_k$, and put $S_k = \sum_{j=1}^k Z_j$. If $\sum_{k=1}^K \mathbb{E}\|Z_k\|^2 \le V$, then for $R > 0$,*

$$\mathbb{P}\left(\max_{1\le k\le K} \|S_k\| > R\right) \le \min\{1, V/R^2\}.$$

If, additionally, the Z_k are independent and $\|Z_k\| \le M$ almost surely for $M > 0$, then for $s > 0$,

$$\mathbb{P}\left(\max_{1\le k\le K} \|S_k\| \ge 2M\sqrt{K} + s\right) \le \exp\left(-\frac{s^2}{2KM^2}\right).$$

Proof. Martingale orthogonality gives $\mathbb{E}\|S_K\|^2 = \sum_k \mathbb{E}\|Z_k\|^2 \le V$. The process $\|S_k\|^2$ is a nonnegative submartingale. Stopping at its first crossing of R^2 (and using bounded-horizon optional sampling) gives $R^2\mathbb{P}(\max_k \|S_k\| > R) \le \mathbb{E}\|S_K\|^2$, proving the first bound.

For the second bound, let $W = \max_k \|S_k\|$. The L^2 maximal inequality gives $\mathbb{E}W^2 \le 4\mathbb{E}\|S_K\|^2 \le 4KM^2$, so $\mathbb{E}W \le 2M\sqrt{K}$. One direct derivation is to stop $\|S_k\|$ at level λ, obtain $\lambda\mathbb{P}(W \ge \lambda) \le \mathbb{E}[\|S_K\|\mathbf{1}_{W\ge\lambda}]$, integrate in λ, and apply Cauchy–Schwarz:

$$\mathbb{E}W^2 \le 2\mathbb{E}[W\|S_K\|] \le 2\sqrt{\mathbb{E}W^2\mathbb{E}\|S_K\|^2}.$$

Changing one independent input Z_j changes W by at most $2M$. The scalar Doob martingale obtained by revealing these inputs therefore has conditional increment ranges of length at most $2M$. Hoeffding's exponential bound for each increment, iterated over the K inputs, gives $\mathbb{P}(W - \mathbb{E}W \ge s) \le \exp(-s^2/(2KM^2))$ Azuma (1967); Vershynin (2018). Combining the two estimates proves the second claim. The mean-size term $2M\sqrt{K}$ is retained rather than identifying a scalar tail bound with an arbitrary vector norm. □

Detailed Proof of Theorem 5. Under the stated local dynamics,

$$\operatorname{dist}(\mathbf{z}_{t_k}, \mathcal{M}) \le \|\mathbf{z}_{t_k} - x_k\| \le \|S_k\| + b_K.$$

Consequently, E_{fail} is contained in $\{\max_k \|S_k\| > \tau - b_K\}$. Conditional centering cannot increase the second moment:

$$\begin{aligned}\mathbb{E}[\|Z_k\|^2 \mid \mathcal{F}_{k-1}] &\le \mathbb{E}[\|X_k\|^2 \mid \mathcal{F}_{k-1}] \\ &= \eta^2\mathbb{E}\left[\|\Delta_k\|^2 \frac{\|\Pi_{N,k}\Delta_k\|^2}{\|\Delta_k\|^2}\,\middle|\,\mathcal{F}_{k-1}\right] \le \eta^2 B^2\zeta^2.\end{aligned}$$

Lemma 1, with $V = K\eta^2B^2\zeta^2$ and $R = \tau - b_K$, proves the proximal bound. For anchor-free directions, the fraction is always at most one, giving the second certificate under the same B and b_K.

If the centered normal increments independently satisfy $\|Z_k\| \le M$ and $\tau - b_K > 2M\sqrt{K}$, the second part of Lemma 1 additionally gives

$$\mathbb{P}(E_{\text{fail}}) \le \exp\left[-\frac{(\tau - b_K - 2M\sqrt{K})^2}{2KM^2}\right].$$

A smaller valid increment bound M improves this exponential certificate. The bounded-increment and independence premises belong to this refinement; the main multi-step certificate above uses conditional second moments and remains applicable without them. □

A.7 Conditional Guarantee: Cross-Attention Surrogate and Manifold Proximity

The geometric objective in §4 is a small normal-energy fraction. Because the full-image normal projector is not directly available during anchor construction, we use cross-attention discrepancy as an optimization surrogate. Its link to normal output variation is formalized by a structural control assumption. The assumption is imposed on the same attention representation, context distribution, and anchor region used in the bound.

Assumption 5 (Attention–Geometry Control)**.** *For the analyzed contexts and anchor representations* $\mathbf{e}_a$*, with* $\Pi_N = \Pi_{N_\mathcal{M}}$*, there exist uniform constants* $L_{\mathrm{CA}} > 0$ *and* $\nu \geq 0$ *such that*

$$\|\Pi_N(\epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_a, t))\| \leq L_{\mathrm{CA}} \sum_{l=1}^{N_L} \|A_l(\mathbf{z}_t, t, \mathbf{e}_u) - A_l(\mathbf{z}_t, t, \mathbf{e}_a)\|_F + \nu. \tag{18}$$

Here N_L *is the number of analyzed attention maps, and* ν *bounds residual normal variation not controlled by their discrepancy. The Frobenius norm is* $\|A\|_F = \sqrt{\sum_i \sum_j |a_{ij}|^2}$*. The constant* L_{CA} *is distinct from the loss* $\mathcal{L}_{\mathrm{CA}}$*;* ν *is retained explicitly in every bound.*

Justification: Cross-attention maps provide spatial text-grounding information Hertz et al. (2022); Tang et al. (2023). Their connection to conditional output variation can be separated into probability-map and value contributions:

$$A_u V_u - A_a V_a = (A_u - A_a)V_u + A_a(V_u - V_a).$$

The map discrepancy controls the first contribution; the bridge condition accounts for value variation and other conditioning paths through L_{CA} and ν. Appendix A.8.5 proves a sufficient local Jacobian condition for this one-way control.

Based on this control assumption, we can formally prove that reducing $\mathcal{L}_{\mathrm{CA}}$ reduces an upper bound on expected normal energy.

Proposition 1 (Normal Drift Bound)**.** *Under Assumption 5, for the unweighted sum-of-squared-map-discrepancies loss in §5.2,*

$$\mathbb{E}_{t,\mathbf{z}_t}[\|\Delta_N^{\mathrm{AB}}\|^2] \leq 2L_{\mathrm{CA}}^2 N_L \mathcal{L}_{\mathrm{CA}}(\mathbf{e}_a) + 2\nu^2. \tag{19}$$

This proposition yields a proximity certificate when the same population loss is small and the prediction difference has nontrivial magnitude. The magnitude condition below complements the semantic regularizer by ensuring a nonzero prediction-difference magnitude.

Theorem 6 (ζ-Proximity via Attention Consistency)**.** *Under Assumption 5, suppose* $\mathcal{L}_{\mathrm{CA}}(\mathbf{e}_a^*) \leq \gamma$ *and* $\|\Delta^{\mathrm{AB}}\|^2 \geq c_{\min} > 0$ *almost surely. If* $2L_{\mathrm{CA}}^2 N_L \gamma + 2\nu^2 < c_{\min}$*, then* $\mathbf{e}_a^*$ *is a* ζ*-proximal anchor for* c_u*, where*

$$\zeta = \sqrt{\frac{2L_{\mathrm{CA}}^2 N_L \gamma + 2\nu^2}{c_{\min}}} < 1. \tag{20}$$

For other positive layer weights, use the corresponding constant in Appendix A.8.4.

Remark: For fixed L_{CA}, ν, and $c_{\min}$, smaller population attention loss tightens the normal-energy certificate. Its limiting level is $\sqrt{2\nu^2/c_{\min}}$, displaying the contribution of residual normal variation. Applying Theorem 2 then gives a tail certificate; applying Theorem 5 additionally uses its per-step and local-dynamics premises. Appendix A.8.8 quantifies transfer to the fitted predictor.

Proof of Proposition 1 and Theorem 6 (Cross-Attention Bridge):

Proof of Proposition 1. By Assumption 5, for each analyzed context,

$$\|\Delta_N^{\mathrm{AB}}\| \leq L_{\mathrm{CA}} \sum_{l=1}^{N_L} \|A_l(u) - A_l(a)\|_F + \nu.$$

Squaring both sides and applying $(x+y)^2 \leq 2x^2 + 2y^2$ gives

$$\|\Delta_N^{\mathrm{AB}}\|^2 \leq 2L_{\mathrm{CA}}^2 \left(\sum_{l=1}^{N_L} \|A_l(u) - A_l(a)\|_F \right)^2 + 2\nu^2.$$

Cauchy–Schwarz Steele (2004) implies $(\sum_l x_l)^2 \leq N_L \sum_l x_l^2$, hence

$$\|\Delta_N^{\mathrm{AB}}\|^2 \leq 2L_{\mathrm{CA}}^2 N_L \sum_{l=1}^{N_L} \|A_l(u) - A_l(a)\|_F^2 + 2\nu^2.$$

Taking expectation over the same context distribution and using the definition of $\mathcal{L}_{\mathrm{CA}}(\mathbf{e}_a)$ proves the proposition. □

Proof of Theorem 6. By the nontrivial-magnitude assumption,

$$\mathbb{E}\left[\frac{\|\Delta_N^{\mathrm{AB}}\|^2}{\|\Delta^{\mathrm{AB}}\|^2}\right] \le \frac{\mathbb{E}\|\Delta_N^{\mathrm{AB}}\|^2}{c_{\min}} \le \frac{2L_{\mathrm{CA}}^2 N_L \mathcal{L}_{\mathrm{CA}}(\mathbf{e}_a^*) + 2\nu^2}{c_{\min}}.$$

Substituting $\mathcal{L}_{\mathrm{CA}}(\mathbf{e}_a^*) \le \gamma$ proves the formula for ζ. The additional strict inequality ensures $\zeta < 1$, as required by Definition 1. The expectation and loss in this calculation use the same context distribution. A held-out finite-sample version is given in Appendix A.8.10. □

A.8 Extensions and Interpretation of the Geometric Results

The following extensions refine the full-image argument through target-concept geometry, stronger concentration under additional premises, and transfer from the surrogate to predictor fitting and measurement. Throughout this section, unsubscripted Π_T and Π_N refer to $\mathcal{M}$ by default.

A.8.1 Reference Geometry and Semantic Erasure

At a clean target reference $\mathbf{z}_u \in \mathcal{M}_u \subseteq \mathcal{M}$, define the additional spaces

$$T_u = T_{\mathbf{z}_u}\mathcal{M}_u, \qquad N_u = T_u^{\perp}, \qquad \Pi_{T_u} = \mathrm{Proj}_{T_u}, \quad \Pi_{N_u} = \mathrm{Proj}_{N_u}.$$

Their dimensions are d_u^* and $d - d_u^*$, providing the target-local counterparts of the full-image spaces $T_{\mathcal{M}}$ and $N_{\mathcal{M}}$ used in the main theorems. Tangent-space inclusion gives

$$T_u \subseteq T_{\mathcal{M}}, \qquad N_u = N_{\mathcal{M}} \oplus (T_{\mathcal{M}} \cap N_u).$$

Consequently, for any direction v,

$$\|\Pi_{N_u} v\|^2 = \|\Pi_N v\|^2 + \|\Pi_{T_{\mathcal{M}} \cap N_u} v\|^2.$$

The full-image normal component measures first-order departure from valid-image geometry. The second component is tangent to $\mathcal{M}$ but normal to $\mathcal{M}_u$ and can include concept-changing motion. The main criterion controls the former while allowing the latter. A target-local normal bound controls both components and therefore upper-bounds first-order departure normal to $\mathcal{M}$. Semantic removal is evaluated through target-conditioned outputs; finite-step structural displacement additionally involves the curvature term in Appendix A.8.9.

For a simple local example, take $\mathcal{M} = \mathbb{R}^2$ and $\mathcal{M}_u$ to be the horizontal axis. The vertical unit vector is normal to $\mathcal{M}_u$ but tangent to $\mathcal{M}$, so its full-image normal energy is zero even though its target-local normal energy is one. This tangent-space example can be realized locally in smooth compact manifolds. This example illustrates the difference between the full-image normal component and the target-local normal component.

Prediction differences and parameter updates. For illustration, consider the squared fitting loss $\frac{1}{2}\|\epsilon_\theta(z, e_u, t) - \epsilon_{\theta^*}(z, e_a, t)\|^2$ at $\theta = \theta^*$, with a frozen reference branch. If J_u is the parameter Jacobian of the target prediction, a gradient step satisfies

$$\delta\theta = -\eta J_u^\top \Delta, \qquad \delta\epsilon_\theta(z, e_u, t) = -\eta J_u J_u^\top \Delta + O(\eta^2).$$

Thus a parameter update generally transforms, rather than copies, the reference difference. Writing $K_u = J_u J_u^\top$, its first-order normal component obeys

$$\|\Pi_N K_u \Delta\| \le \|\Pi_N K_u \Pi_N\|_{\mathrm{op}} \|\Delta_N\| + \|\Pi_N K_u \Pi_T\|_{\mathrm{op}} \|\Delta_T\|.$$

The first term transfers the normal component of the reference signal; the second records tangent-to-normal mixing by the parameter Jacobian. These terms specify the sufficient quantities for controlling an optimization step. For the unsquared norm in $\mathcal{L}_{\mathrm{rm}}$, a nonzero single-context residual gives the same leading direction divided by $\|\Delta\|$, so the scalar step factor changes. Appendix A.8.8 supplies a complementary result stated directly in terms of the final fitting error, independent of the optimization path.

A.8.2 Normal Energy under a Weaker Local-Alignment Condition

Let v be any unit prediction-difference direction, with Π_T and Π_N relative to $\mathcal{M}$, and let $a \in [0,1]$, $\delta > 0$, and $\beta(\delta) \in [0,1]$. If

$$\mathbb{P}(\|\Pi_T v\|^2 > a + \delta) \le \beta(\delta),$$

then orthogonal decomposition gives the exact implication

$$\mathbb{P}(\|\Pi_N v\|^2 \ge 1 - a - \delta) \ge 1 - \beta(\delta).$$

Likewise, $\mathbb{E}\|\Pi_T v\|^2 \le a$ implies $\mathbb{E}\|\Pi_N v\|^2 \ge 1 - a$. Neither implication needs isotropy. The proof is simply $\|\Pi_N v\|^2 = 1 - \|\Pi_T v\|^2$. Theorem 1 supplies a particular dimension-dependent tail under its stronger spherical premise. For structured learned directions, a and β retain their alignment-specific values. Empirical estimates of mean alignment are related to population quantities in Appendices A.8.6 and A.8.10.

A.8.3 Exponential Refinement for Proximal Anchors

The mean-based statement in Theorem 2 admits an exponential refinement when the full-image normal-energy ratio concentrates across contexts. Let $q = \|\Delta_N^{\mathrm{PA}}\|^2 / \|\Delta^{\mathrm{PA}}\|^2$ on the context distribution in Definition 1.

Assumption 6 (Exponential Concentration of Proximal Normal Energy)**.** *For the analyzed proximal anchor, the centered normal-energy ratio satisfies*

$$\mathbb{E}\exp\{\lambda(q - \mathbb{E}q)\} \le \exp(2\lambda^2/d), \qquad \lambda \ge 0.$$

This condition concerns fluctuation about the mean. Definition 1 separately controls the mean itself. It is an additional premise of the exponential refinement.

Exponential suppression. Under Definition 1 and Assumption 6, for every $\delta > 0$,

$$\mathbb{P}(q \ge \zeta^2 + \delta) \le \exp(-d\delta^2/8).$$

Consequently, the more conservative threshold that retains the dimensional term also satisfies

$$\mathbb{P}\left(q \ge \zeta^2 + \frac{d^*}{d} + \delta\right) \le \exp(-d\delta^2/8).$$

Proof. Definition 1 gives $\mathbb{E}q \le \zeta^2$. For $\lambda > 0$, exponential Markov inequality and Assumption 6 yield

$$\begin{aligned}\mathbb{P}(q \ge \zeta^2 + \delta) &\le \mathbb{P}(q - \mathbb{E}q \ge \delta) \\ &\le e^{-\lambda\delta}\mathbb{E}e^{\lambda(q-\mathbb{E}q)} \le \exp(-\lambda\delta + 2\lambda^2/d).\end{aligned}$$

The exponent is minimized at $\lambda = d\delta/4$, yielding $-d\delta^2/8$. The event with threshold $\zeta^2 + d^*/d + \delta$ is a subset of the event just bounded. If a threshold exceeds one, the corresponding event is empty, consistent with the same inequality. □

A more general fluctuation scale. For a condition $\mathbb{E}e^{\lambda(q-\mathbb{E}q)} \le e^{v_q\lambda^2/2}$ with $v_q > 0$, the same optimization gives $\mathbb{P}(q \ge \zeta^2 + \delta) \le e^{-\delta^2/(2v_q)}$. Assumption 6 specializes this scale to $v_q = 4/d$. A sufficient mathematical example is $q = m^{-1}\sum_{j=1}^{m} Y_j$ for independent $Y_j \in [0,1]$ and $m \ge d/16$: the bounded-variable MGF bound gives

$$\log \mathbb{E}e^{\lambda(q-\mathbb{E}q)} \le \frac{\lambda^2}{8m} \le \frac{2\lambda^2}{d}.$$

For completeness, for a random variable Y supported on an interval of length one, the second derivative of $\log \mathbb{E}e^{\lambda Y}$ is its variance under exponential tilting, bounded by $1/4$. Integrating twice after centering gives $\log \mathbb{E}e^{\lambda(Y-\mathbb{E}Y)} \le \lambda^2/8$; independence then gives the displayed averaged bound. This construction supplies a sufficient condition for Assumption 6. The mean-based bound in Theorem 2 follows from Definition 1; the absolute bound in Theorem 4 additionally uses the stated magnitude condition $\|\Delta^{\mathrm{PA}}\| \le B$.

A.8.4 Attention Normalization and the Implemented Surrogate

The sum form in Proposition 1 uses fixed maps A_l. More generally, for positive weights w_l and the same map family, define

$$\mathcal{L}_{\mathrm{CA}}^{(w)} = \mathbb{E} \sum_{l=1}^{N_L} w_l \|A_l(u) - A_l(a)\|_F^2.$$

Weighted Cauchy–Schwarz gives

$$\left(\sum_l x_l \right)^2 \leq \left(\sum_l w_l^{-1} \right) \left(\sum_l w_l x_l^2 \right).$$

Thus, under Assumption 5 for those same maps,

$$\mathbb{E}\|\Delta_N^{\mathrm{AB}}\|^2 \leq 2L_{\mathrm{CA}}^2 \left(\sum_l w_l^{-1} \right) \mathcal{L}_{\mathrm{CA}}^{(w)} + 2\nu^2.$$

Theorem 6 applies with $N_L\gamma$ replaced by $(\sum_l w_l^{-1})\gamma_w$, where γ_w bounds this weighted loss. Setting $w_l = 1$ recovers the original proposition.

For the span-restricted, head-averaged maps $\overline{A}_l$ described in Appendix B, the implemented mean-of-layer-MSE objective corresponds, for a fixed shape, to $A_l = \overline{A}_l$ and $w_l = 1/(N_L n_l)$, with n_l the number of entries in each compared map. Equivalently, rescaling each map by $\sqrt{w_l}$ expresses this same loss as an unweighted sum, provided the bridge constants are defined for the rescaled maps. This normalization uses the actual selected, head-averaged maps in the bridge condition. Shape-dependent weights are handled by a uniform bound or a context-wise rescaled formulation. The loss thresholds and bridge constants consequently refer to the same representation.

A.8.5 A Sufficient Local Condition for the Attention Bridge

We now derive a local sufficient condition for the attention–geometry bridge. Fix a context and the full-image projector Π_N. Let $F(e)$ stack the actual attention representations used for comparison, and let $G(e) = \Pi_N \epsilon_{\theta^*}(\mathbf{z}_t, e, t)$. Suppose F and G are twice continuously differentiable near e_u, and their Jacobians satisfy

$$\ker DF(e_u) \subseteq \ker DG(e_u).$$

Then there is a linear map H on the range of $DF(e_u)$ such that $DG(e_u) = HDF(e_u)$. Indeed, define $H(DF(e_u)h) = DG(e_u)h$; the kernel inclusion makes this well-defined. Extend H by zero on the orthogonal complement and let $L = \|H\|_{\mathrm{op}}$.

If the operator norms of the second derivatives of F and G are bounded by M_F and M_G on a radius-r ball, Taylor's theorem gives, for $\|h\| \leq r$,

$$\begin{aligned} F(e_u + h) - F(e_u) &= DF(e_u)h + R_F, & \|R_F\| &\leq \tfrac{1}{2} M_F \|h\|^2, \\ G(e_u + h) - G(e_u) &= DG(e_u)h + R_G, & \|R_G\| &\leq \tfrac{1}{2} M_G \|h\|^2. \end{aligned}$$

Therefore,

$$\|G(e_u + h) - G(e_u)\| \leq L\|F(e_u + h) - F(e_u)\| + \tfrac{1}{2}(LM_F + M_G)r^2.$$

For stacked maps, their Euclidean discrepancy is at most the sum of individual Frobenius discrepancies, yielding Assumption 5 on that neighborhood with $\nu = \frac{1}{2}(LM_F + M_G)r^2$. Uniform application across contexts requires uniform constants and neighborhoods. The condition says that a conditioning direction invisible to the chosen attention representation must also be invisible to the first-order normal output. For an application to a selected map family, the Jacobian condition and the uniform derivative bounds specify the required local model properties. The radius r is the neighborhood on which those properties hold; an anchor region contained in that neighborhood inherits the bound.

A.8.6 Connection to PCA and Downstream Measurements

The reported PCA experiment uses a target-conditioned clean-latent bank to estimate the target-local tangent space. Formally, let $P_{u,i} = \Pi_{T_{\mathbf{z}_{u,i}}\mathcal{M}_u}$ be the true target-local tangent projector and $\widehat{P}_{u,i} = Q_i Q_i^\top$ its empirical estimate. For a unit prediction-difference direction $v_i = \Delta_i/\|\Delta_i\|$, define

$$q_{u,i} = v_i^\top (I - P_{u,i}) v_i, \qquad q_{\mathcal{M},i} = \|\Pi_{N_{\mathcal{M},i}} v_i\|^2, \qquad e_i = \|P_{u,i} - \widehat{P}_{u,i}\|_{\mathrm{op}}.$$

The deterministic projector-error inequality is

$$\left| q_{u,i} - v_i^\top (I - \widehat{P}_{u,i}) v_i \right| \le e_i.$$

With the experimental stabilizer $\varepsilon > 0$, write

$$\widehat{q}_i^{\,\varepsilon} = \frac{\|(I - \widehat{P}_{u,i})\Delta_i\|^2}{\|\Delta_i\|^2 + \varepsilon}.$$

If $\|\Delta_i\|^2 \ge c_{\min} > 0$, the difference from the unstabilized estimated ratio is at most $\varepsilon/(c_{\min} + \varepsilon)$. Combining this bound with Appendix A.8.1, sample by sample,

$$q_{\mathcal{M},i} \le q_{u,i} \le \widehat{q}_i^{\,\varepsilon} + e_i + \frac{\varepsilon}{c_{\min} + \varepsilon}.$$

Averaging preserves this deterministic error bound, which tightly bounds the true target-local alignment provided the estimation error e_i and the stabilization ratio are sufficiently small. Appendix A.8.10 incorporates an additional sampling term for an independent held-out bank. This error-bound formulation equally applies to a full-image estimator given a corresponding reference projector and estimation error over $\mathcal{M}$.

Table 14 reports target-local estimated normal-energy ratios 0.880, 0.540, 0.350, and 0.080 for empty, manual, Stage-I, and optimized-anchor directions; the random control is 0.980. These observations quantify alignment with the estimated target-local space. The projector-error inequality above and Proposition 5 relate these measurements to a population upper bound on $\mathbb{E}q_{\mathcal{M}}$, with explicit projector-estimation, numerical-stabilization, and sampling corrections. Table 16 evaluates the sensitivity of this diagnostic across explained-variance thresholds of 80%, 90%, and 95%, with respective ranks 260, 327, and 363.

The empirical evaluations jointly assess concept removal and utility preservation across multiple complementary facets. The two-stage ablations and fixed-anchor studies analyze method-specific trade-offs, while the target-residual audit (Appendix C.9) quantifies output-level target dominance against held-out alternatives. Additionally, the sequential experiments evaluate cumulative removal and utility preservation over consecutive edits. Finally, if two native predictor parameterizations are related at a fixed context by $f_t(z, e) = a_t \epsilon_\theta(z, e, t) + b_t(z)$ with a condition-independent scalar $a_t \neq 0$, their conditional differences yield identical normal-energy fractions. This geometric invariance holds for any valid reference geometry within a given model context. Extending the framework to distinct pretrained backbones correspondingly relies on their specific prediction functions and intrinsic geometric properties.

A.8.7 Connecting Anchor-Free and Manual-Anchor References

The two method families differ in the reference used to form the conditioning signal. This subsection makes that connection explicit before applying the directional bounds. Fix a context $(\mathbf{z}_t, t)$ and the same full-image normal projector Π_N for every reference. Let $\mathbf{e}_a^*$ be any chosen comparison reference, and write

$$\begin{aligned} \Delta^{\mathrm{PA}} &= \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_u, t) - \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_a^*, t), \\ r_b &= \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_a^*, t) - \epsilon_{\theta^*}(\mathbf{z}_t, \mathbf{e}_b, t), \qquad b \in \{\emptyset, a\}. \end{aligned}$$

When the comparison reference is proximal, Δ^{PA} has the meaning used in Definition 1. The algebra holds for any fixed comparison reference:

$$\Delta^{\mathrm{AF}} = \Delta^{\mathrm{PA}} + r_\emptyset, \qquad \Delta^{\mathrm{AB}} = \Delta^{\mathrm{PA}} + r_a.$$

Thus empty-prompt and manual-anchor removal share the same algebraic structure, but incur different reference mismatches.

Proposition 2 (Normal mismatch decomposition)**.** *For either reference b, linearity and the reverse triangle inequality give*

$$\|\Pi_N \Delta_b\| \geq \big(\|\Pi_N r_b\| - \|\Pi_N \Delta^{\mathrm{PA}}\|\big)_+, \qquad (x)_+ = \max\{x, 0\}.$$

If the compared signals are square-integrable over the same context distribution, the reverse triangle inequality in L^2 similarly gives

$$\big(\mathbb{E}\|\Pi_N \Delta_b\|^2\big)^{1/2} \geq \Big[\big(\mathbb{E}\|\Pi_N r_b\|^2\big)^{1/2} - \big(\mathbb{E}\|\Pi_N \Delta^{\mathrm{PA}}\|^2\big)^{1/2}\Big]_+ .$$

Proof. Use $\Pi_N \Delta_b = \Pi_N \Delta^{\mathrm{PA}} + \Pi_N r_b$ and rearrange the triangle inequality, first in the Euclidean norm and then in the norm of square-integrable random vectors. Nonnegativity yields the positive part. □

The bound quantifies when reference mismatch dominates the normal component of the comparison direction. It applies to both empty-prompt and manual references, with their respective mismatch terms. The weaker local-alignment result in Appendix A.8.2 and the isotropic specialization in Theorem 1 quantify complementary aspects of this same geometry.

Anchor-free removal. For exact conditional and marginal scores and a discrete target event c_u, Bayes' rule gives

$$\nabla_{\mathbf{z}} \log p_t(\mathbf{z} \mid c_u) - \nabla_{\mathbf{z}} \log p_t(\mathbf{z}) = \nabla_{\mathbf{z}} \log p_t(c_u \mid \mathbf{z}).$$

In noise-prediction units, the corresponding ideal difference is $-\sigma_t \nabla_{\mathbf{z}} \log p_t(c_u \mid \mathbf{z})$. It is a target-discriminative signal relative to the marginal distribution, rather than a signal toward a specified replacement concept. This explains why learning a valid unconditional image distribution does not by itself impose the required tangent alignment on the conditional-minus-unconditional difference. The identity assumes that the empty branch represents that marginal score; approximation errors remain as defined in Appendix A.2.

Manual-anchor removal. An exact target-to-anchor score difference instead has the form

$$\nabla_{\mathbf{z}} \log p_t(\mathbf{z} \mid c_u) - \nabla_{\mathbf{z}} \log p_t(\mathbf{z} \mid c_a) = \nabla_{\mathbf{z}} \log \frac{p_t(\mathbf{z} \mid c_u)}{p_t(\mathbf{z} \mid c_a)}.$$

Specifying c_a selects a replacement, but semantic similarity alone imposes no bound on the normal component of this likelihood-ratio gradient. A local construction illustrates the missing implication. Let $\mathbf{t}$ and $\mathbf{n}$ be unit tangent and normal vectors, and let the smooth predictor be $f(e_1, e_2) = e_1\mathbf{t} + e_2\mathbf{n}$ with target representation $(0, 0)$. References $(h, 0)$ and $(0, h)$ have the same arbitrarily small Euclidean distance $h > 0$ to the target. Their prediction differences nevertheless have normal-energy fractions zero and one, respectively. Thus proximity in a representation space and proximity of the induced direction are distinct geometric quantities. This example motivates assessing the prediction difference itself, as in Definition 1.

Interpretation of reference-induced bias. Reference-induced bias denotes dependence of the removal–preservation trade-off on the chosen reference. Statistical mean displacement and directional dispersion enter its normal-energy analysis through the following decomposition. For a random vector $Y = \Pi_N r_b$ in a fixed ambient coordinate system,

$$\mathbb{E}\|Y\|^2 = \|\mathbb{E}Y\|^2 + \mathbb{E}\|Y - \mathbb{E}Y\|^2.$$

Large normal energy can therefore arise from either a mean offset or dispersion. Directional normal energy therefore captures both contributions. The fixed-anchor study measures reference sensitivity through output removal and preservation, while the main directional result quantifies the normal-energy consequence of the stated alignment model.

A.8.8 From an Anchor Target to the Edited Predictor

The removal objective trains the target-conditioned prediction to match the anchor-conditioned reference. This provides a direct connection between the geometry of the reference difference and the final edited prediction.

Proposition 3 (Transfer through prediction fitting). *On the common target-conditioned context distribution, with Π_N defined on $\mathcal{M}$, let*

$$f_u = \epsilon_{\theta^*}(z_t, e_u, t), \qquad f_a = \epsilon_{\theta^*}(z_t, e_a^*, t), \qquad f_{\mathrm{edit}} = \epsilon_{\theta_u}(z_t, e_u, t),$$

and define $\Delta = f_u - f_a$, $r_{\mathrm{fit}} = f_{\mathrm{edit}} - f_a$, and $D = f_{\mathrm{edit}} - f_u$. Then

$$D = -\Delta + r_{\mathrm{fit}}, \qquad \|\Pi_N D\| \le \|\Pi_N \Delta\| + \|\Pi_N r_{\mathrm{fit}}\|.$$

If the anchor is ζ-proximal, $\|\Delta\| \le B$ almost surely, and $\mathbb{E}\|r_{\mathrm{fit}}\|^2 \le \varepsilon_{\mathrm{fit}}^2$, then

$$\mathbb{E}\|\Pi_N D\|^2 \le 2B^2\zeta^2 + 2\varepsilon_{\mathrm{fit}}^2.$$

If instead a pointwise fitting bound $\|r_{\mathrm{fit}}\| \le \varepsilon_{\mathrm{fit}}$ is available, then for $\delta > 0$,

$$\mathbb{P}\Big(\|\Pi_N D\| > B\sqrt{\zeta^2 + \delta} + \varepsilon_{\mathrm{fit}}\Big) \le \frac{\zeta^2}{\zeta^2 + \delta}.$$

Under Assumption 6, the last probability is also at most $\exp(-d\delta^2/8)$.

Proof. The identity follows by adding and subtracting f_a. Orthogonal projection and the triangle inequality give the pointwise estimate. Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$\mathbb{E}\|\Pi_N D\|^2 \le 2\mathbb{E}\|\Pi_N \Delta\|^2 + 2\mathbb{E}\|r_{\mathrm{fit}}\|^2 \le 2B^2\mathbb{E}\frac{\|\Pi_N \Delta\|^2}{\|\Delta\|^2} + 2\varepsilon_{\mathrm{fit}}^2.$$

Definition 1 gives the stated mean-square bound. With pointwise fitting control, the event in the proposition implies $\|\Pi_N \Delta\| > B\sqrt{\zeta^2 + \delta}$ and hence $q > \zeta^2 + \delta$. Apply Theorem 2, or the exponential refinement when its additional premise holds. □

This result applies to any optimization procedure reaching the stated fitting error on the analyzed contexts. It accommodates the norm-based $\mathcal{L}_{\mathrm{rm}}$ in Section 5: the pointwise fitting error is that norm, while the mean-square result uses its second moment. The regularized objective and its stopping rule determine the fitting error achieved in a particular run. Thus reference alignment and fitting accuracy contribute separately and explicitly to the geometry of the edited predictor.

A.8.9 Local Image-Geometry Control with a Curvature Remainder

A local graph representation converts full-image normal control into a bound on distance from the valid-image manifold, with an explicit second-order curvature term.

Proposition 4 (Local structural displacement). *At $x \in \mathcal{M}$, write a neighborhood of the smooth embedded manifold $\mathcal{M}$ as*

$$\{x + v + g_x(v) : v \in T_x\mathcal{M},\ \|v\| < r\},$$

where $g_x(v) \in N_x\mathcal{M}$, $g_x(0) = 0$, $Dg_x(0) = 0$, and $\|D^2 g_x(v)\|_{\mathrm{op}} \le C_g$ on the neighborhood. For any displacement h with $\|\Pi_{T_x\mathcal{M}} h\| < r$,

$$\mathrm{dist}(x + h, \mathcal{M}) \le \|\Pi_{N_x\mathcal{M}} h\| + \frac{C_g}{2}\|\Pi_{T_x\mathcal{M}} h\|^2.$$

For the context-wise application, take $x = \mathbf{z}_u$, let r and C_g be uniform over the analyzed contexts, and put $h = \eta D$ as in Proposition 3. If $\|D\| \le B_D$ almost surely, $\eta B_D < r$, and the pointwise fitting bound holds, then with probability at least $1 - \zeta^2/(\zeta^2 + \delta)$,

$$\mathrm{dist}(x + \eta D, \mathcal{M}) \le \eta\big(B\sqrt{\zeta^2 + \delta} + \varepsilon_{\mathrm{fit}}\big) + \frac{C_g}{2}\eta^2 B_D^2,$$

The step-size condition keeps the displacement in the graph domain. The exponential confidence from Appendix A.8.3 applies when its premise holds.

Proof. Take $v = \Pi_{T_x\mathcal{M}} h$ and compare $x + h$ to the manifold point $x + v + g_x(v)$. Their distance is at most $\|\Pi_{N_x\mathcal{M}} h\| + \|g_x(v)\|$. Taylor's theorem gives $\|g_x(v)\| \le C_g\|v\|^2/2$. This proves the first assertion with the full-image normal term. At $x = \mathbf{z}_u$, substituting $h = \eta D$ and applying Proposition 3 proves the second assertion. At a target reference, $\|\Pi_{N_x\mathcal{M}} h\| \le \|\Pi_{N_u} h\|$ also gives the conservative target-local version recorded in Appendix A.8.11. □

The estimate separates reference misalignment, prediction-fitting error, and a second-order curvature remainder. It concerns a local displacement in the latent geometry. It supplies a sufficient structural-control relation; the downstream CLIP and FID measurements evaluate semantic removal and generation utility on the actual sampled outputs.

A.8.10 Finite-Sample Certificates on Held-Out Contexts

To connect empirical diagnostics to population quantities, fix the optimized anchor, any shared reference bank, and the rule for estimating the local projector before drawing an independent validation bank. For this certificate, each projector estimate is then a fixed function of its own validation context and that pre-fixed reference bank. Let the resulting n validation contexts be independent and identically distributed under the context distribution of Definition 1. The constants and distribution in this subsection refer to that validation protocol. Its projectors $P_{u,i}$ and $\widehat{P}_{u,i}$ are target-local as in Appendix A.8.6; $q_{\mathcal{M}}$ is the full-image ratio in Definition 1, and q_u is its stronger target-local counterpart.

Proposition 5 (Held-out proximity and surrogate control)**.** *Suppose $\|\Delta_i\|^2 \geq c_{\min} > 0$ and $\|P_{u,i} - \widehat{P}_{u,i}\|_{\mathrm{op}} \leq e_{\max}$ almost surely. With stabilized measurements $\widehat{q}_i^{\varepsilon}$ as in Appendix A.8.6, for $\alpha \in (0,1)$, with probability at least $1-\alpha$ over the validation bank,*

$$\mathbb{E}q_{\mathcal{M}} \leq \mathbb{E}q_u \leq \frac{1}{n}\sum_{i=1}^{n} \widehat{q}_i^{\varepsilon} + e_{\max} + \frac{\varepsilon}{c_{\min}+\varepsilon} + \sqrt{\frac{\log(1/\alpha)}{2n}}.$$

If the per-context surrogate loss ℓ_i satisfies $0 \leq \ell_i \leq M_\ell$ almost surely for a fixed $M_\ell > 0$, its population loss likewise satisfies, with probability at least $1-\alpha$,

$$\mathcal{L}_{\mathrm{CA}}(e_a^*) \leq \widehat{\mathcal{L}}_{\mathrm{CA}}(e_a^*) + M_\ell\sqrt{\frac{\log(1/\alpha)}{2n}}.$$

Proof. Appendix A.8.6 gives $q_{\mathcal{M},i} \leq q_{u,i} \leq \widehat{q}_i^{\varepsilon} + e_{\max} + \varepsilon/(c_{\min}+\varepsilon)$ pointwise. Taking population expectations and applying the bounded-variable MGF argument from Appendix A.8.3 to the independent variables $\widehat{q}_i^{\varepsilon} \in [0,1]$ gives $\mathbb{P}(\mathbb{E}\widehat{q}^{\varepsilon} - n^{-1}\sum_i \widehat{q}_i^{\varepsilon} \geq t) \leq e^{-2nt^2}$. Choose $t = \sqrt{\log(1/\alpha)/(2n)}$. Apply the same argument to $\ell_i/M_\ell \in [0,1]$ for the second bound. The two assertions hold separately at level $1-\alpha$; replacing α by $\alpha/2$ in each yields a joint statement by the union bound. □

The upper confidence bound in the second inequality can be used as γ in Theorem 6 under its stated structural and magnitude conditions. For application to the reported PCA means, the projector-error bounds and validation sampling protocol determine the correction terms. Independence is evaluated conditional on the pre-fixed anchor, reference bank, and projector-estimation rule. Shared random contexts or reuse of validation data in fitting the projector require the corresponding dependence to be accounted for when defining n.

A.8.11 Target-Concept Specialization and Local Certificates

Theorems 1–6 use the full image manifold $\mathcal{M}$. This subsection specializes their projection arguments to $\mathcal{M}_u$ and records the additional premises required for the target-local counterparts. The context law remains the same target-conditioned law throughout.

Two reference geometries. At $\mathbf{z}_u \in \mathcal{M}_u$, use T_u, N_u from Appendix A.8.1, and define, for a nonzero difference Δ,

$$\Delta_{N,u} = \Pi_{N_u}\Delta, \qquad q_u = \frac{\|\Delta_{N,u}\|^2}{\|\Delta\|^2}, \qquad q_{\mathcal{M}} = \frac{\|\Pi_N\Delta\|^2}{\|\Delta\|^2}, \qquad q_{\mathrm{sem}} = \frac{\|\Pi_{T_{\mathcal{M}}\cap N_u}\Delta\|^2}{\|\Delta\|^2}.$$

The orthogonal decomposition gives the exact relation

$$q_u = q_{\mathcal{M}} + q_{\mathrm{sem}}, \qquad 0 \leq q_{\mathcal{M}} \leq q_u \leq 1.$$

The full-image criterion permits a nonzero q_{sem}, which measures the energy fraction in $T_{\mathcal{M}} \cap N_u$. This subspace can include concept-changing directions. Target-local proximity additionally controls this component. See Appendix A.8.1 for an example with $q_{\mathrm{sem}} > 0$ and $q_{\mathcal{M}} = 0$.

Directional concentration on $\mathcal{M}_u$. Suppose $0 < d_u^* < d$ and the analyzed normalized difference v is uniform on the unit sphere conditional on T_u. Then the target-local counterpart of Theorem 1 is

$$\mathbb{P}\left(q_u > 1 - \frac{d_u^*}{d} - \delta\right) \geq 1 - \exp(-d\delta^2/8), \qquad \delta \in [0,1).$$

Proof. Condition on T_u and repeat the chi-squared projection calculation in Appendix A.6, now with tangent rank $m = d_u^*$. The squared tangent projection is distributed as $U/(U+V)$ for independent $U \sim \chi^2_{d_u^*}$ and $V \sim \chi^2_{d-d_u^*}$. The same exponential-moment bound and the identity $q_u = 1 - \|\Pi_{T_u} v\|^2$ yield the assertion. □

The conditional law is stated for the projector used in each bound.[3] For structured directions, Appendix A.8.2 applies with Π_T, Π_N replaced by Π_{T_u}, Π_{N_u} and with the target-local alignment parameters.

Target-local proximity and upper-bound transfer. Call an anchor target-locally ζ_u-proximal when its difference is nonzero almost surely and $\mathbb{E}q_u \le \zeta_u^2 < 1$. Its full-image ratio automatically satisfies $\mathbb{E}q_{\mathcal{M}} \le \zeta_u^2$, so it also satisfies Definition 1 with the same upper-bound parameter. Markov's inequality and, where applicable, $\|\Delta\| \le B$ give

$$\mathbb{P}(q_u \ge \zeta_u^2 + \delta) \le \frac{\zeta_u^2}{\zeta_u^2 + \delta}, \qquad \mathbb{P}\Big(\|\eta\Delta_{N,u}\| > \eta B\sqrt{\zeta_u^2 + \delta}\Big) \le \frac{\zeta_u^2}{\zeta_u^2 + \delta}.$$

These are the target-local versions of Theorems 2 and 4. Each upper bound also controls the corresponding full-image event by $q_{\mathcal{M}} \le q_u$ or by projection-norm dominance.

Conversely, a full-image certificate $\mathbb{E}q_{\mathcal{M}} \le \zeta^2$ extends to a target-local certificate if the additional semantic-tangent energy satisfies $\mathbb{E}q_{\mathrm{sem}} \le \omega_u^2$ and $\zeta^2 + \omega_u^2 < 1$: one may then take $\zeta_u^2 = \zeta^2 + \omega_u^2$. This follows by taking expectations in the exact decomposition.

If the target-local ratio separately obeys

$$\mathbb{E}e^{\lambda(q_u - \mathbb{E}q_u)} \le e^{2\lambda^2/d}, \qquad \lambda \ge 0,$$

then Appendix A.8.3 also gives $\mathbb{P}(q_u \ge \zeta_u^2 + \delta) \le e^{-d\delta^2/8}$. The same target-derived tail upper-bounds the full-image event by event inclusion. The concentration premises for $q_{\mathcal{M}}$ and q_u are stated separately.

Target-normal Gaussian specialization. For a target-local counterpart of Theorem 3, impose the joint Gaussian error model from Appendix A.2 on the larger normal space N_u, now of dimension $n_u = d - d_u^*$. Denote its marginal scale by $\sigma_{\mathrm{est},u}$ and correlation by $\rho^{(u)}_{c_u,\emptyset} < 1$, with $\rho^{(u)}_{c_u,\emptyset} \ge -1$. Conditional on the fixed context, let

$$\Delta^{\mathrm{AF}}_{N,u} = \Pi_{N_u} S + \sigma_{\mathrm{eff},u} G_u, \qquad G_u \sim \mathcal{N}(0, I_{n_u}), \qquad \sigma^2_{\mathrm{eff},u} = 2\sigma^2_{\mathrm{est},u}(1 - \rho^{(u)}_{c_u,\emptyset}).$$

The negative-exponential calculation in Appendix A.6 yields

$$\mathbb{P}\Big(\|\eta\Delta^{\mathrm{AF}}_{N,u}\| \ge \eta\sigma_{\mathrm{eff},u}\sqrt{(d - d_u^*)(1-\delta)}\Big) \ge 1 - e^{-(d-d_u^*)\delta^2/8}, \qquad 0 < \delta < 1.$$

The full-image Gaussian model specifies errors only in $N_{\mathcal{M}}$. The target-local version additionally specifies errors in $T_{\mathcal{M}} \cap N_u$, with separate parameters. Likewise, the signal–error upper-bound refinement in Appendix A.6 applies with n_u and the target-local Gaussian parameters.

Target-local attention and fitting certificates. Suppose the target-local bridge holds on the same context distribution and anchor region:

$$\|\Pi_{N_u}\Delta^{\mathrm{AB}}\| \le L_{\mathrm{CA},u}\sum_{l=1}^{N_L}\|A_l(u) - A_l(a)\|_F + \nu_u, \qquad L_{\mathrm{CA},u} > 0, \quad \nu_u \ge 0.$$

Squaring, applying Cauchy–Schwarz, and taking expectations as in Proposition 1 give

$$\mathbb{E}\|\Delta^{\mathrm{AB}}_{N,u}\|^2 \le 2L^2_{\mathrm{CA},u} N_L \mathcal{L}_{\mathrm{CA}}(e_a) + 2\nu_u^2.$$

[3] A spherical law conditional on both reference geometries is sufficient for both versions; conditioning on a larger random collection of geometric information requires a separate directional premise.

If the population loss is at most γ, $\|\Delta^{\mathrm{AB}}\|^2 \geq c_{\min} > 0$ almost surely, and the numerator below is smaller than $c_{\min}$, then

$$\zeta_u^2 = \frac{2L_{\mathrm{CA},u}^2 N_L \gamma + 2\nu_u^2}{c_{\min}} < 1$$

is a target-local proximity certificate. Its residual level is controlled by ν_u. This target-local bridge implies the full-image Assumption 5 with the same constants via projection-norm dominance. A target-local certificate can also be obtained from a full-image certificate combined with a bound on $\mathbb{E}q_{\mathrm{sem}}$. Appendix A.8.5 supplies a local sufficient condition for the target-local bridge by using $G_u(e) = \Pi_{N_u}\epsilon_{\theta^*}(z_t, e, t)$ and the corresponding target-normal Jacobian and derivative bounds.

The exact reference-mismatch identities in Proposition 2 remain valid with Π_{N_u}. For fitting transfer, $D = -\Delta + r_{\mathrm{fit}}$ still holds, and target-local proximity gives

$$\mathbb{E}\|\Pi_{N_u} D\|^2 \leq 2B^2\zeta_u^2 + 2\varepsilon_{\mathrm{fit}}^2$$

under the same magnitude and fitting-error conditions as Proposition 3. At $x \in \mathcal{M}_u$, Proposition 4 and normal-space inclusion also preserve the conservative local image-distance bound

$$\mathrm{dist}(x + h, \mathcal{M}) \leq \|\Pi_{N_u} h\| + \frac{C_g}{2}\|\Pi_{T_x\mathcal{M}} h\|^2$$

on its graph domain. The curvature constant here belongs to $\mathcal{M}$, because the distance is measured to $\mathcal{M}$.

Target-local multi-step stability. A process-level specialization uses predictable reference points $x_k^u \in \mathcal{M}_u$ and projectors $\Pi_{N_u,k} = \Pi_{N_{x_k^u}\mathcal{M}_u}$. Define

$$X_k^u = \eta\Pi_{N_u,k}\Delta_k, \quad Z_k^u = X_k^u - \mathbb{E}[X_k^u \mid \mathcal{F}_{k-1}], \quad S_k^u = \sum_{j=1}^{k} Z_j^u.$$

Assume

$$\mathbf{z}_{t_k} = x_k^u + S_k^u + \mathbf{r}_k^u, \qquad \max_{k\leq K}\|\mathbf{r}_k^u\| \leq b_{K,u} < \tau_u,$$

$0 < \|\Delta_k\| \leq B$ almost surely, and $\mathbb{E}[\|\Pi_{N_u,k}\Delta_k\|^2/\|\Delta_k\|^2 \mid \mathcal{F}_{k-1}] \leq \zeta_u^2$ at every step. The proof of Theorem 5, with these centers and projectors, gives

$$\mathbb{P}\left(\max_{k\leq K}\mathrm{dist}(\mathbf{z}_{t_k}, \mathcal{M}_u) > \tau_u\right) \leq \min\left\{1, \frac{K\eta^2B^2\zeta_u^2}{(\tau_u - b_{K,u})^2}\right\}.$$

Indeed, conditional centering bounds the total martingale variance by $K\eta^2B^2\zeta_u^2$, and Lemma 1 controls the maximum at radius $\tau_u - b_{K,u}$. An anchor-free process with the same target-local magnitude and remainder premises has the analogous bound with ζ_u^2 replaced by one. When τ_u is selected using a tubular neighborhood, the reach and curvature conditions are imposed on $\mathcal{M}_u$ directly.

For this same process and tolerance, $\mathrm{dist}(z, \mathcal{M}) \leq \mathrm{dist}(z, \mathcal{M}_u)$, so its target-local certificate also bounds departure from $\mathcal{M}$. The main full-image process in Theorem 5 allows reference points anywhere on $\mathcal{M}$. These certificates control geometric displacement; semantic erasure is assessed by the conditioned outputs.

Role of the reported diagnostic. The empirical PCA bank continues to estimate the target-local ratio. Thus, the target-local PCA measurements connect to the full-image bound through the projector-error and sampling terms in Appendices A.8.6 and A.8.10.

B Implementation Details

Experimental setting and reporting. Unless otherwise specified, experiments use Mickey Mouse, Stable Diffusion v1.4 (SD1.4), a candidate budget of $K = 20$, and GPT-5.6 Sol as the Stage-I proposer. Quantitative model-editing results summarize five runs of anchor optimization and downstream editing. A score is displayed as its raw mean with relative standard deviation (RSD) in

the subscript, where $\mathrm{RSD} = 100s/|\bar{m}|$ and s is the across-run standard deviation. The percentage subscript is not a confidence interval or the standard deviation of the normalized gain. When an experiment reuses a baseline or default-setting result, both its mean and RSD are reused; these entries do not represent additional independent runs. Gains are computed from the reported means, and method-level gains are averaged without including the standalone method. The same ten generation seeds are used across compared methods: 1234, 101, 67153984, 23, 6, 23512, 13142, 293293, 3223, and 6125. The target-residual audit uses the single-checkpoint protocol in Appendix C.9.

Stage-I candidate construction. Stage I starts from $N = 10$ LLM-generated sentences containing the target concept c_u. After masking the target occurrence, we collect the returned substitute concepts, rank them by their occurrence counts, and retain up to K candidates to form C_0. Sentences and returned proposals vary with the target and generation randomness; the LLM is not used during Stage II, model editing, or image generation. We apply K-means to the candidate and target representations, discard candidates in the target-associated cluster, and average the remaining set C_1 as in Eq. (4). Candidate generation is repeated when filtering leaves insufficient cluster diversity.

Candidate-length alignment and aggregation. Construction operates on token-lookup embeddings before positional embeddings and the text transformer. Let L_u be the target-span length and $E_c \in \mathbb{R}^{L_c \times 768}$ a candidate's lookup sequence. When length alignment is required, we mean-pool the candidate tokens and replicate the pooled vector over the target span:

$$\bar{e}_c = \frac{1}{L_c} \sum_{j=1}^{L_c} (E_c)_j, \qquad (\widetilde{E}_c)_k = \bar{e}_c, \quad k = 1, \ldots, L_u.$$

The aligned candidates have a common shape $L_u \times 768$ and can be averaged elementwise to obtain E_a^0. This is a sequence-level summary for initialization, not an average of contextualized prompt embeddings. Stage II optimizes the continuous entries of this tensor; the initialization does not require all span positions to remain identical after optimization. For SD1.4, a single anchor has shape $[1, L_u, 768]$.

Full-prompt replacement and EOT re-encoding. For a complete prompt y, let $E(y)$ denote its native lookup sequence. We find complete occurrences of the target token span and replace their lookup vectors by the aligned anchor:

$$\widetilde{E}_j(y, E_a) = \begin{cases} (E_a)_k, & j \text{ is position } k \text{ in a matched target span,} \\ E_j(y), & \text{otherwise,} \end{cases} \qquad H_a(y) = T_{\text{text}}\big(\widetilde{E}(y, E_a) + P; M\big).$$

Here P and M are the native positional embeddings and attention masks. Prompt length, special-token positions, and the EOT lookup vector are preserved. Running the full frozen text transformer on the modified sequence recomputes all contextualized states, including EOT. Thus EOT re-encoding refers to its output hidden state; it does not zero the EOT input or append the old contextualized EOT. Non-target lookup vectors stay unchanged, although their contextualized states can change with the substituted span.

Stage-II objective. The text encoder and diffusion backbone are frozen while gradients propagate to the inserted continuous anchor. Target and anchor branches use the same latent state and timestep. For SD1.4, we select the target-span columns of post-softmax cross-attention maps, average over attention heads, and average the per-layer mean-squared discrepancy over eight selected layers. With head-averaged maps $\bar{A}_l$ and n_l compared entries in layer l,

$$\mathcal{L}_{\text{CA}} = \mathbb{E}\left[\frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{\|\bar{A}_l(e_u) - \bar{A}_l(e_a)\|_F^2}{n_l} \right].$$

Appendix A.8.4 relates this normalization to Eq. (5). Euclidean distances are computed on flattened target-length lookup tensors. The retained cluster centroids define the target-centered distance range in Eq. (6), and its soft hinge penalty uses $\lambda = 1.0$. There is no additional hard-projection step.

Model editing. After Stage II, E_a^* is fixed. Each removal prompt is re-encoded with the optimized span to obtain its anchor-conditioned representation. The frozen reference prediction uses this representation, while the student prediction uses the ordinary target-conditioned prompt. The student is trained with the removal and preservation terms in Section 5. Re-encoding constructs the conditioning for the current prompt; it does not jointly re-optimize the anchor and student. In plug-in comparisons, the existing baseline objective and trainable parameter subset are retained, and the optimized anchor supplies the replacement conditioning reference.

Evaluation and runtime. CLIP scores use ViT-L/14 with $100\cos(f_I(x), f_T(y))$, without the learned logit-scale multiplier. CLIP(U) measures full target-containing prompt alignment, while CLIP(R) uses the non-target retain prompts. FID compares Inception-v3 features of approximately 30,000 generated images with the COCO train2017 reference set; the pretrained SD1.4 FID is 15.58 under this protocol. Appendix C.10 times Stage-II anchor optimization plus model editing until early stopping. Stage-I querying and aggregation are reported separately, and no LLM call is made during image generation.

C Supplementary Experimental Results

In this appendix, we provide extensive supplementary experiments and in-depth analyses to further validate the effectiveness, robustness, and efficiency of AutoAnchor .

To demonstrate the generalizability of our method, we first present complete quantitative results for the two-stage ablation study (§C.4)and evaluate AutoAnchor on expanded target concepts (§C.1). We then explore the scalability of our approach by testing on modern DiT-based architectures, including SD3.5-M and FLUX.2-4B (§C.2), and by assessing its stability in sequential multi-concept unlearning scenarios (§C.3).

Furthermore, we investigate the framework's behavior and sensitivity through a dedicated case study on bias unlearning with fixed manual anchors (§C.7). We also present comprehensive ablation studies verifying the robustness of AutoAnchor across various LLM proposers (§C.6) and hyperparameter candidate budgets (§C.5). Finally, we examine the computational efficiency and runtime overhead of our method (§C.10), and provide a local geometric diagnostic (§C.8) alongside a held-out target-residual audit (§C.9) to geometrically and empirically verify the manifold-proximal properties of our synthesized anchors.

All experimental results presented in the main text and the appendices were obtained as the average of five independent measurements using one NVIDIA A800 GPU with 80GB memory.

C.1 Additional Results on Expanded Target Concepts

We extend the SD1.4 evaluation to seven additional targets: Porsche, Alaskan Malamute, Obama identity, Einstein identity, Picasso style, Grumpy, and Violence. Tables 3, 4, and 5 retain the object/identity, style, and attribute/content groupings. All gains use the pretrained-relative definitions in Table 1. Each Average row is the arithmetic mean of seven method-level gains and excludes standalone AutoAnchor .

Enhanced Concept Removal. Standalone AutoAnchor achieves the lowest or second-lowest CLIP(U) for all seven targets. Across the 49 baseline–target pairs, augmentation lowers CLIP(U) in 46 cases. The mean additional target-alignment reductions are 36.47% for Porsche, 45.54% for Grumpy, 28.38% for Picasso, 274.84% for Alaskan Malamute, 10.20% for Obama, 67.39% for Violence, and 11.02% for Einstein. The remaining regressions are tuneSD-AF on Obama and FMN and tuneSD-AB on Einstein. The large normalized mean for Alaskan Malamute is sensitive to the FMN denominator: its original CLIP(U) is only 0.10 below the pretrained reference. The corresponding raw scores are therefore important when interpreting this percentage.

Improved Model Utility. Augmentation improves CLIP(R) in 46 of 49 pairs and lowers FID in all 49. The mean reductions in retained-alignment degradation, in the order Porsche, Grumpy, Picasso, Alaskan Malamute, Obama, Violence, and Einstein, are 37.49%, 15.77%, 24.71%, 19.53%, 60.33%, 46.66%, and 57.97%. The corresponding mean excess-FID reductions are 53.11%, 49.94%, 28.62%, 20.88%, 30.62%, 25.36%, and 26.89%. Standalone AutoAnchor has the lowest FID on five targets

Table 3: **Additional Object and Identity Targets on SD1.4.** We follow the setting in Table 1 and report unlearning performance for object and identity targets: Porsche, Alaskan Malamute, Obama identity, and Einstein identity.

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Porsche *Pretrained:* $U_0 = 35.33,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$29.50_{\pm 1.1\%}$	$26.68_{\pm 0.7\%}$	+48.37	$31.42_{\pm 1.4\%}$	$33.86_{\pm 0.9\%}$	+56.74	$17.61_{\pm 0.6\%}$	$16.21_{\pm 1.2\%}$	+68.97
tuneSD-AF	$34.28_{\pm 0.9\%}$	$34.18_{\pm 1.3\%}$	+9.52	$32.51_{\pm 1.0\%}$	$32.45_{\pm 1.1\%}$	-1.87	$21.16_{\pm 1.4\%}$	$19.25_{\pm 0.8\%}$	+34.23
ResAlign	$31.69_{\pm 1.2\%}$	$31.61_{\pm 0.6\%}$	+2.20	$32.01_{\pm 0.8\%}$	$33.28_{\pm 1.3\%}$	+34.23	$18.47_{\pm 1.0\%}$	$16.67_{\pm 1.1\%}$	+62.28
CCRT	$28.46_{\pm 1.3\%}$	$\underline{26.33}_{\pm 1.0\%}$	+31.00	$31.47_{\pm 0.7\%}$	$32.09_{\pm 1.4\%}$	+14.59	$16.64_{\pm 0.9\%}$	$\underline{16.10}_{\pm 1.2\%}$	+50.94
ConAbl	$30.79_{\pm 0.8\%}$	$30.50_{\pm 1.1\%}$	+6.39	$31.57_{\pm 1.2\%}$	$31.89_{\pm 0.6\%}$	+7.71	$17.61_{\pm 1.3\%}$	$16.29_{\pm 1.0\%}$	+65.02
FMN	$33.87_{\pm 1.0\%}$	$32.43_{\pm 1.4\%}$	+98.63	$32.16_{\pm 0.9\%}$	$\underline{34.79}_{\pm 1.2\%}$	+73.88	$21.00_{\pm 0.7\%}$	$19.83_{\pm 1.1\%}$	+21.59
tuneSD-AB	$31.29_{\pm 1.4\%}$	$28.90_{\pm 0.8\%}$	+59.16	$31.87_{\pm 1.1\%}$	$\mathbf{34.84}_{\pm 1.0\%}$	+77.14	$18.62_{\pm 0.6\%}$	$16.53_{\pm 1.3\%}$	+68.75
Average	–	–	+36.47	–	–	+37.49	–	–	+53.11
AutoAnchor	–	$\mathbf{26.32}_{\pm 1.2\%}$	–	–	$34.03_{\pm 0.9\%}$	–	–	$\mathbf{16.03}_{\pm 1.1\%}$	–
Alaskan Malamute *Pretrained:* $U_0 = 35.05,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$27.45_{\pm 1.3\%}$	$27.34_{\pm 0.9\%}$	+1.45	$30.48_{\pm 1.1\%}$	$32.35_{\pm 1.4\%}$	+35.69	$18.79_{\pm 0.7\%}$	$18.02_{\pm 1.2\%}$	+23.99
tuneSD-AF	$31.77_{\pm 1.0\%}$	$30.74_{\pm 1.3\%}$	+31.40	$33.42_{\pm 0.8\%}$	$33.46_{\pm 1.1\%}$	+1.74	$19.72_{\pm 1.4\%}$	$18.81_{\pm 0.6\%}$	+21.98
ResAlign	$31.60_{\pm 1.2\%}$	$29.91_{\pm 1.0\%}$	+48.99	$33.57_{\pm 0.9\%}$	$33.79_{\pm 1.3\%}$	+10.23	$24.33_{\pm 0.8\%}$	$22.95_{\pm 1.1\%}$	+15.77
CCRT	$25.20_{\pm 1.1\%}$	$\underline{25.13}_{\pm 0.7\%}$	+0.71	$31.95_{\pm 1.4\%}$	$33.04_{\pm 1.0\%}$	+28.91	$18.21_{\pm 1.2\%}$	$\underline{17.29}_{\pm 0.9\%}$	+34.98
ConAbl	$31.12_{\pm 0.6\%}$	$30.03_{\pm 1.2\%}$	+27.74	$32.75_{\pm 1.3\%}$	$33.12_{\pm 1.1\%}$	+12.46	$20.63_{\pm 1.0\%}$	$19.66_{\pm 0.8\%}$	+19.21
FMN	$34.95_{\pm 1.4\%}$	$33.16_{\pm 1.0\%}$	+1790.00	$35.19_{\pm 0.9\%}$	$\mathbf{35.39}_{\pm 1.2\%}$	+37.74	$22.90_{\pm 1.1\%}$	$21.89_{\pm 1.3\%}$	+13.80
tuneSD-AB	$30.43_{\pm 1.2\%}$	$29.34_{\pm 0.8\%}$	+23.59	$33.30_{\pm 1.0\%}$	$33.54_{\pm 1.4\%}$	+9.92	$21.41_{\pm 0.7\%}$	$20.45_{\pm 1.1\%}$	+16.47
Average	–	–	+274.84	–	–	+19.53	–	–	+20.88
AutoAnchor	–	$\mathbf{25.11}_{\pm 1.1\%}$	–	–	$\underline{35.36}_{\pm 1.3\%}$	–	–	$\mathbf{17.23}_{\pm 1.0\%}$	–
Obama Identity *Pretrained:* $U_0 = 35.72,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$29.67_{\pm 1.1\%}$	$\underline{28.75}_{\pm 0.9\%}$	+15.21	$32.34_{\pm 1.2\%}$	$33.32_{\pm 1.0\%}$	+28.99	$18.77_{\pm 0.8\%}$	$17.13_{\pm 1.2\%}$	+51.41
tuneSD-AF	$34.28_{\pm 1.0\%}$	$34.30_{\pm 1.3\%}$	-1.39	$33.69_{\pm 0.9\%}$	$35.01_{\pm 1.1\%}$	+65.02	$22.50_{\pm 1.3\%}$	$21.62_{\pm 0.7\%}$	+12.72
ResAlign	$33.85_{\pm 1.2\%}$	$33.08_{\pm 1.0\%}$	+41.18	$33.19_{\pm 1.1\%}$	$34.96_{\pm 1.4\%}$	+69.96	$19.66_{\pm 0.9\%}$	$18.78_{\pm 1.1\%}$	+21.57
CCRT	$29.08_{\pm 0.8\%}$	$28.96_{\pm 1.2\%}$	+1.81	$32.53_{\pm 1.0\%}$	$33.76_{\pm 0.9\%}$	+38.56	$17.25_{\pm 1.1\%}$	$\underline{16.89}_{\pm 1.0\%}$	+21.56
ConAbl	$31.61_{\pm 1.3\%}$	$31.45_{\pm 0.8\%}$	+3.89	$32.09_{\pm 0.7\%}$	$34.33_{\pm 1.2\%}$	+61.71	$18.34_{\pm 1.0\%}$	$17.52_{\pm 1.3\%}$	+29.71
FMN	$33.63_{\pm 1.0\%}$	$33.57_{\pm 1.1\%}$	+2.87	$32.27_{\pm 1.2\%}$	$34.89_{\pm 0.8\%}$	+75.94	$19.69_{\pm 1.1\%}$	$18.15_{\pm 1.0\%}$	+37.47
tuneSD-AB	$32.67_{\pm 0.9\%}$	$32.43_{\pm 1.2\%}$	+7.87	$32.31_{\pm 1.0\%}$	$\underline{35.11}_{\pm 1.3\%}$	+82.11	$18.74_{\pm 1.2\%}$	$17.48_{\pm 0.9\%}$	+39.87
Average	–	–	+10.20	–	–	+60.33	–	–	+30.62
AutoAnchor	–	$\mathbf{22.09}_{\pm 1.1\%}$	–	–	$\mathbf{35.39}_{\pm 1.2\%}$	–	–	$\mathbf{16.84}_{\pm 1.0\%}$	–
Einstein Identity *Pretrained:* $U_0 = 35.29,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$29.51_{\pm 1.2\%}$	$\underline{28.88}_{\pm 0.9\%}$	+10.90	$32.16_{\pm 1.0\%}$	$33.13_{\pm 1.3\%}$	+27.25	$18.44_{\pm 0.8\%}$	$17.50_{\pm 1.1\%}$	+32.87
tuneSD-AF	$33.43_{\pm 1.1\%}$	$33.15_{\pm 0.7\%}$	+15.05	$33.87_{\pm 1.2\%}$	$\underline{35.10}_{\pm 1.0\%}$	+66.49	$20.48_{\pm 1.0\%}$	$19.22_{\pm 1.3\%}$	+25.71
ResAlign	$33.71_{\pm 0.9\%}$	$32.89_{\pm 1.2\%}$	+51.90	$33.43_{\pm 1.1\%}$	$34.81_{\pm 0.8\%}$	+60.26	$20.35_{\pm 1.2\%}$	$19.48_{\pm 1.0\%}$	+18.24
CCRT	$29.20_{\pm 1.0\%}$	$29.05_{\pm 1.3\%}$	+2.46	$32.36_{\pm 0.9\%}$	$33.52_{\pm 1.1\%}$	+34.52	$18.35_{\pm 1.1\%}$	$17.53_{\pm 0.9\%}$	+29.60
ConAbl	$31.76_{\pm 1.2\%}$	$31.29_{\pm 1.0\%}$	+13.31	$32.26_{\pm 1.3\%}$	$34.51_{\pm 1.1\%}$	+65.03	$18.10_{\pm 0.8\%}$	$17.96_{\pm 1.2\%}$	+5.56
FMN	$33.47_{\pm 1.1\%}$	$33.73_{\pm 0.9\%}$	-14.29	$32.02_{\pm 1.0\%}$	$34.74_{\pm 1.3\%}$	+73.51	$19.81_{\pm 1.2\%}$	$18.44_{\pm 1.0\%}$	+32.39
tuneSD-AB	$32.54_{\pm 0.9\%}$	$32.60_{\pm 1.1\%}$	-2.18	$32.15_{\pm 1.0\%}$	$34.96_{\pm 1.2\%}$	+78.71	$18.43_{\pm 1.1\%}$	$\mathbf{17.18}_{\pm 1.0\%}$	+43.86
Average	–	–	+11.02	–	–	+57.97	–	–	+26.89
AutoAnchor	–	$\mathbf{22.21}_{\pm 1.0\%}$	–	–	$\mathbf{35.12}_{\pm 1.2\%}$	–	–	$\underline{17.22}_{\pm 1.1\%}$	–

and the second-lowest on Einstein; for Violence, ESD and ConAbl augmented with AutoAnchor have lower FID. These results support improved removal–preservation trade-offs across the tested targets, rather than uniform dominance on every metric.

C.2 Evaluation on Modern DiT-based Architectures

To demonstrate the generalizability and scalability of AutoAnchor, we extend our evaluation beyond traditional U-Net architectures to modern state-of-the-art Diffusion Transformer (DiT) models. Specifically, we evaluate the unlearning performance on Stable Diffusion 3.5 Medium (SD3.5-M) and FLUX.2-4B across three diverse target concepts: Mickey Mouse (object), Van Gogh style (style),

Table 4: **Additional Image Style Unlearning Concepts on SD1.4.** We follow the setting in Table 1 and report unlearning performance for the Picasso style concept.

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Picasso Style *Pretrained:* $U_0 = 35.84,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$28.40_{\pm 0.6\%}$	$26.82_{\pm 1.4\%}$	+21.24	$33.29_{\pm 1.2\%}$	$33.88_{\pm 0.9\%}$	+24.28	$17.66_{\pm 1.0\%}$	$\underline{16.30}_{\pm 1.1\%}$	+65.38
tuneSD-AF	$34.04_{\pm 1.3\%}$	$33.09_{\pm 1.0\%}$	+52.78	$\mathbf{34.67}_{\pm 0.8\%}$	$34.57_{\pm 1.4\%}$	-9.52	$18.21_{\pm 1.2\%}$	$17.49_{\pm 0.7\%}$	+27.38
ResAlign	$31.89_{\pm 1.1\%}$	$30.27_{\pm 1.3\%}$	+41.01	$33.58_{\pm 0.9\%}$	$34.52_{\pm 1.0\%}$	+43.93	$20.29_{\pm 0.6\%}$	$19.35_{\pm 1.2\%}$	+19.96
CCRT	$24.64_{\pm 1.0\%}$	$\mathbf{23.61}_{\pm 1.4\%}$	+9.20	$31.38_{\pm 1.3\%}$	$33.45_{\pm 1.1\%}$	+47.70	$16.94_{\pm 0.8\%}$	$16.66_{\pm 1.0\%}$	+20.59
ConAbl	$32.09_{\pm 1.2\%}$	$30.84_{\pm 0.9\%}$	+33.33	$33.89_{\pm 1.0\%}$	$34.50_{\pm 1.3\%}$	+33.33	$20.46_{\pm 1.1\%}$	$19.36_{\pm 0.7\%}$	+22.54
FMN	$33.63_{\pm 0.7\%}$	$33.23_{\pm 1.2\%}$	+18.10	$33.67_{\pm 1.4\%}$	$34.10_{\pm 1.0\%}$	+20.98	$21.33_{\pm 1.3\%}$	$19.93_{\pm 1.1\%}$	+24.35
tuneSD-AB	$31.75_{\pm 1.1\%}$	$30.81_{\pm 0.8\%}$	+22.98	$33.68_{\pm 1.2\%}$	$33.93_{\pm 0.6\%}$	+12.25	$21.00_{\pm 1.0\%}$	$19.91_{\pm 1.4\%}$	+20.11
Average	–	–	+28.38	–	–	+24.71	–	–	+28.62
AutoAnchor	–	$\underline{23.81}_{\pm 1.2\%}$	–	–	$\underline{34.63}_{\pm 1.0\%}$	–	–	$\mathbf{16.15}_{\pm 1.1\%}$	–

Table 5: **Additional Attribute and Content Targets on SD1.4.** We follow the setting in Table 1 and report unlearning performance for the Grumpy attribute and Violence content target.

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Grumpy *Pretrained:* $U_0 = 35.90,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$31.71_{\pm 0.7\%}$	$30.13_{\pm 1.3\%}$	+37.71	$33.60_{\pm 1.0\%}$	$34.11_{\pm 1.2\%}$	+24.06	$16.79_{\pm 1.4\%}$	$16.40_{\pm 0.8\%}$	+32.23
tuneSD-AF	$32.55_{\pm 1.1\%}$	$31.22_{\pm 0.6\%}$	+39.70	$33.64_{\pm 1.3\%}$	$33.71_{\pm 1.0\%}$	+3.37	$17.45_{\pm 0.9\%}$	$16.82_{\pm 1.2\%}$	+33.69
ResAlign	$31.82_{\pm 1.2\%}$	$29.91_{\pm 1.1\%}$	+46.81	$33.79_{\pm 0.8\%}$	$34.06_{\pm 1.4\%}$	+13.99	$16.72_{\pm 1.0\%}$	$16.19_{\pm 0.7\%}$	+46.49
CCRT	$31.33_{\pm 0.9\%}$	$\underline{29.64}_{\pm 1.2\%}$	+36.98	$32.97_{\pm 1.1\%}$	$33.85_{\pm 0.6\%}$	+32.00	$17.17_{\pm 1.3\%}$	$16.28_{\pm 1.0\%}$	+55.97
ConAbl	$32.18_{\pm 0.8\%}$	$30.59_{\pm 1.4\%}$	+42.74	$33.76_{\pm 1.2\%}$	$34.21_{\pm 0.9\%}$	+22.96	$18.18_{\pm 1.1\%}$	$16.90_{\pm 1.3\%}$	+49.23
FMN	$32.50_{\pm 1.0\%}$	$30.58_{\pm 0.7\%}$	+56.47	$\underline{34.23}_{\pm 1.4\%}$	$34.22_{\pm 1.1\%}$	-0.67	$17.18_{\pm 0.6\%}$	$\underline{16.09}_{\pm 1.2\%}$	+68.13
tuneSD-AB	$32.32_{\pm 1.3\%}$	$30.23_{\pm 1.0\%}$	+58.38	$33.75_{\pm 0.8\%}$	$34.04_{\pm 1.3\%}$	+14.72	$17.24_{\pm 1.1\%}$	$16.18_{\pm 0.9\%}$	+63.86
Average	–	–	+45.54	–	–	+15.77	–	–	+49.94
AutoAnchor	–	$\mathbf{29.62}_{\pm 1.1\%}$	–	–	$\mathbf{34.33}_{\pm 1.0\%}$	–	–	$\mathbf{16.07}_{\pm 1.2\%}$	–
Violence *Pretrained:* $U_0 = 35.88,\ R_0 = 35.72,\ F_0 = 15.58.$									
ESD	$30.16_{\pm 1.7\%}$	$25.39_{\pm 0.6\%}$	+83.39	$31.52_{\pm 1.1\%}$	$33.20_{\pm 1.9\%}$	+40.00	$17.06_{\pm 0.8\%}$	$\mathbf{16.52}_{\pm 1.4\%}$	+36.49
tuneSD-AF	$31.33_{\pm 1.5\%}$	$30.15_{\pm 1.8\%}$	+25.93	$33.25_{\pm 1.8\%}$	$\underline{34.76}_{\pm 2.1\%}$	+61.13	$20.43_{\pm 1.1\%}$	$19.28_{\pm 0.9\%}$	+23.71
ResAlign	$32.02_{\pm 1.5\%}$	$27.63_{\pm 0.8\%}$	+113.73	$33.71_{\pm 1.0\%}$	$34.62_{\pm 1.6\%}$	+45.27	$20.66_{\pm 0.6\%}$	$19.49_{\pm 2.1\%}$	+23.03
CCRT	$26.92_{\pm 1.2\%}$	$\mathbf{21.02}_{\pm 1.8\%}$	+65.85	$29.81_{\pm 0.7\%}$	$31.53_{\pm 1.0\%}$	+29.10	$17.82_{\pm 1.9\%}$	$17.33_{\pm 0.5\%}$	+21.88
ConAbl	$30.97_{\pm 1.4\%}$	$25.16_{\pm 0.6\%}$	+118.33	$33.00_{\pm 1.7\%}$	$34.55_{\pm 1.1\%}$	+56.99	$18.34_{\pm 0.9\%}$	$\underline{17.04}_{\pm 1.3\%}$	+47.10
FMN	$31.85_{\pm 0.9\%}$	$31.16_{\pm 1.1\%}$	+17.12	$33.56_{\pm 1.1\%}$	$34.34_{\pm 1.3\%}$	+36.11	$22.25_{\pm 1.3\%}$	$21.83_{\pm 1.6\%}$	+6.30
tuneSD-AB	$30.01_{\pm 0.9\%}$	$27.23_{\pm 1.3\%}$	+47.36	$31.46_{\pm 1.4\%}$	$33.93_{\pm 0.6\%}$	+57.98	$21.99_{\pm 1.2\%}$	$20.77_{\pm 1.7\%}$	+19.03
Average	–	–	+67.39	–	–	+46.66	–	–	+25.36
AutoAnchor	–	$\underline{21.18}_{\pm 1.1\%}$	–	–	$\mathbf{34.81}_{\pm 1.4\%}$	–	–	$17.26_{\pm 0.6\%}$	–

and Nude (content). Tables 6 and 7 present the quantitative results for SD3.5-M and FLUX.2-4B, respectively.

Enhanced Concept Removal. The results consistently confirm that AutoAnchor maintains superior removal efficacy on modern DiT backbones. When operating as a standalone method, AutoAnchor consistently achieves the best or second-best CLIP(U) scores across both models and all three concepts. Furthermore, when integrated into existing baselines, AutoAnchor-augmented methods demonstrate substantial improvements in concept removal. On SD3.5-M, the average CLIP(U) improvements are 41.61%, 46.31%, and 38.21% for the three concepts. Similarly, on FLUX.2-4B, the integration of AutoAnchor yields impressive average CLIP(U) enhancements of 44.22%, 53.10%, and 43.65%. These results support the transfer of automatic anchor construction to the two tested transformer-based diffusion backbones.

Improved Model Utility. In addition to enhanced concept removal, AutoAnchor improves the removal–utility trade-off on the tested backbones. Standalone AutoAnchor consistently preserves the highest or second-highest text-image alignment (CLIP(R)) and image quality (FID) compared

Table 6: **Evaluation on SD3.5 Medium.** We show the experimental results using modern DiT architectures, SD3.5 Medium, following the same setting as in Table 1 except for the backbone and its pretrained references. Subscripts report RSD (%).

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
SD3.5-M: Mickey Mouse *Pretrained:* $U_0 = 39.11,\ R_0 = 39.15,\ F_0 = 11.03$.									
ESD	$28.43_{\pm 1.7\%}$	$24.66_{\pm 0.6\%}$	+35.30	$32.48_{\pm 1.1\%}$	$34.15_{\pm 1.9\%}$	+25.04	$16.48_{\pm 0.8\%}$	$14.90_{\pm 1.4\%}$	+28.99
tuneSD-AF	$34.66_{\pm 0.5\%}$	$32.43_{\pm 2.0\%}$	+50.11	$36.08_{\pm 1.3\%}$	**38.41**$_{\pm 0.9\%}$	+75.90	$18.32_{\pm 1.2\%}$	$15.11_{\pm 0.7\%}$	+44.03
ResAlign	$29.74_{\pm 1.5\%}$	$27.49_{\pm 0.8\%}$	+24.01	$34.54_{\pm 1.0\%}$	$36.52_{\pm 1.6\%}$	+42.95	$17.14_{\pm 0.6\%}$	$15.90_{\pm 2.1\%}$	+20.29
CCRT	$25.79_{\pm 1.2\%}$	$\underline{21.76}_{\pm 1.8\%}$	+30.26	$31.35_{\pm 0.7\%}$	$31.94_{\pm 1.0\%}$	+7.56	$14.56_{\pm 1.9\%}$	$\underline{13.38}_{\pm 0.5\%}$	+33.43
ConAbl	$30.80_{\pm 1.4\%}$	$27.27_{\pm 0.6\%}$	+42.48	$32.45_{\pm 1.7\%}$	$35.80_{\pm 1.1\%}$	+50.00	$16.66_{\pm 0.9\%}$	$15.07_{\pm 1.3\%}$	+28.24
tuneSD-AB	$31.32_{\pm 0.8\%}$	$26.06_{\pm 1.5\%}$	+67.52	$32.92_{\pm 1.2\%}$	$36.58_{\pm 0.7\%}$	+58.75	$17.23_{\pm 1.0\%}$	$15.76_{\pm 1.8\%}$	+23.71
Average	–	–	+41.61	–	–	+43.37	–	–	+29.78
AutoAnchor	–	**19.76**$_{\pm 1.1\%}$	–	–	$\underline{37.90}_{\pm 1.4\%}$	–	–	**13.29**$_{\pm 0.6\%}$	–
SD3.5-M: Van Gogh Style *Pretrained:* $U_0 = 39.25,\ R_0 = 39.15,\ F_0 = 11.03$.									
ESD	$28.83_{\pm 1.6\%}$	$25.21_{\pm 0.9\%}$	+34.74	$32.99_{\pm 1.0\%}$	$34.97_{\pm 2.0\%}$	+32.14	$16.91_{\pm 0.7\%}$	$15.12_{\pm 1.5\%}$	+30.44
tuneSD-AF	$35.48_{\pm 1.3\%}$	$33.11_{\pm 0.5\%}$	+62.86	$36.89_{\pm 0.8\%}$	**39.48**$_{\pm 1.8\%}$	+114.60	$18.74_{\pm 1.1\%}$	$15.45_{\pm 0.6\%}$	+42.67
ResAlign	$30.42_{\pm 1.9\%}$	$28.12_{\pm 1.2\%}$	+26.05	$35.24_{\pm 1.1\%}$	$37.36_{\pm 0.7\%}$	+54.22	$17.61_{\pm 1.4\%}$	$16.68_{\pm 0.9\%}$	+14.13
CCRT	$26.42_{\pm 1.0\%}$	$\underline{22.28}_{\pm 1.6\%}$	+32.27	$31.99_{\pm 1.3\%}$	$32.67_{\pm 2.1\%}$	+9.50	$14.89_{\pm 0.5\%}$	**13.42**$_{\pm 1.0\%}$	+38.08
ConAbl	$31.55_{\pm 0.6\%}$	$27.94_{\pm 1.4\%}$	+46.88	$33.16_{\pm 1.7\%}$	$36.63_{\pm 0.8\%}$	+57.93	$17.02_{\pm 1.2\%}$	$15.42_{\pm 1.9\%}$	+26.71
tuneSD-AB	$32.08_{\pm 1.5\%}$	$26.70_{\pm 0.7\%}$	+75.03	$33.71_{\pm 0.9\%}$	$37.43_{\pm 1.3\%}$	+68.38	$17.66_{\pm 1.8\%}$	$16.12_{\pm 0.6\%}$	+23.23
Average	–	–	+46.31	–	–	+56.13	–	–	+29.21
AutoAnchor	–	**20.25**$_{\pm 1.2\%}$	–	–	$\underline{38.72}_{\pm 0.9\%}$	–	–	$\underline{13.61}_{\pm 1.7\%}$	–
SD3.5-M: Nude *Pretrained:* $U_0 = 39.18,\ R_0 = 39.15,\ F_0 = 11.03$.									
ESD	$28.15_{\pm 0.8\%}$	$24.43_{\pm 1.4\%}$	+33.73	$32.66_{\pm 1.9\%}$	$34.32_{\pm 0.6\%}$	+25.58	$16.41_{\pm 1.1\%}$	$14.81_{\pm 1.6\%}$	+29.74
tuneSD-AF	$34.44_{\pm 1.3\%}$	$32.67_{\pm 0.9\%}$	+37.34	$35.91_{\pm 0.7\%}$	**38.63**$_{\pm 1.5\%}$	+83.95	$18.25_{\pm 1.0\%}$	$15.03_{\pm 2.0\%}$	+44.60
ResAlign	$29.91_{\pm 1.7\%}$	$27.62_{\pm 0.5\%}$	+24.70	$34.41_{\pm 1.2\%}$	$36.68_{\pm 1.8\%}$	+47.89	$17.08_{\pm 0.6\%}$	$15.98_{\pm 1.2\%}$	+18.18
CCRT	$25.95_{\pm 1.4\%}$	$\underline{21.63}_{\pm 1.0\%}$	+32.65	$31.19_{\pm 0.8\%}$	$32.12_{\pm 1.3\%}$	+11.68	$14.48_{\pm 1.9\%}$	$\underline{13.47}_{\pm 0.7\%}$	+29.28
ConAbl	$30.61_{\pm 1.1\%}$	$27.08_{\pm 1.6\%}$	+41.19	$32.63_{\pm 0.6\%}$	$35.62_{\pm 1.1\%}$	+45.86	$16.57_{\pm 1.5\%}$	$15.14_{\pm 0.8\%}$	+25.81
tuneSD-AB	$31.10_{\pm 0.9\%}$	$26.28_{\pm 1.2\%}$	+59.65	$33.08_{\pm 1.7\%}$	$36.41_{\pm 0.6\%}$	+54.86	$17.34_{\pm 1.2\%}$	$15.69_{\pm 1.8\%}$	+26.15
Average	–	–	+38.21	–	–	+44.97	–	–	+28.96
AutoAnchor	–	**19.62**$_{\pm 1.3\%}$	–	–	$\underline{38.07}_{\pm 1.0\%}$	–	–	**13.38**$_{\pm 1.5\%}$	–

to the baselines. More importantly, augmenting existing methods with AutoAnchor systematically improves their non-target utility. For SD3.5-M, the augmented baselines achieve average CLIP(R) improvements of up to 56.13% and consistent FID reductions of approximately 29%. For FLUX.2-4B, we observe average CLIP(R) gains of up to 44.49% and average FID improvements ranging from 38.74% to 40.66%. Across all 36 method–backbone–target pairs, augmentation lowers CLIP(U) and FID; CLIP(R) improves in 33 pairs, with the three exceptions occurring for ConAbl on FLUX.2-4B. These comparisons measure downstream removal and preservation on the tested architectures.

C.3 Sequential Multi-Concept Unlearning

We evaluate three sequential edits of the same SD1.4 model: **Mickey Mouse** → **Van Gogh style** → **Nude**. At each checkpoint, CLIP(U) covers all targets erased so far and uses the measured cumulative pretrained reference. Table 8 reports removal, retained alignment, and image quality. Step 1 reuses the main-text Mickey Mouse results, including their RSDs.

Cumulative concept removal. Mean additional target-alignment reduction increases from 21.69% in Step 1 to 41.26% in Step 2 and 77.49% in Step 3. At the final checkpoint, every augmented baseline has lower cumulative CLIP(U) than its unaugmented counterpart; the normalized gains for CCRT and ConAbl are 160.25% and 95.92%. Standalone AutoAnchor has cumulative CLIP(U) of 18.61, 20.14, and 20.56 across the three steps. These are cumulative prompt-set measurements rather than individual-concept rebound scores.

Table 7: **Evaluation on FLUX.2 4B.** We show the experimental results using modern DiT architectures, FLUX.2 4B, following the same setting as in Table 1 except for the backbone and its pretrained references. Subscripts report RSD (%).

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
FLUX.2-4B: Mickey Mouse *Pretrained:* $U_0 = 39.35,\ R_0 = 39.30,\ F_0 = 10.98$.									
ESD	$30.38_{\pm 1.2\%}$	$26.40_{\pm 0.8\%}$	+44.37	$33.52_{\pm 1.7\%}$	$36.19_{\pm 0.6\%}$	+46.19	$17.56_{\pm 1.4\%}$	$15.67_{\pm 1.9\%}$	+28.72
tuneSD-AF	$35.24_{\pm 0.9\%}$	$31.07_{\pm 1.1\%}$	+101.46	$35.23_{\pm 1.5\%}$	**38.13**$_{\pm 0.7\%}$	+71.25	$16.91_{\pm 1.3\%}$	$13.53_{\pm 0.5\%}$	+57.00
ResAlign	$31.89_{\pm 1.8\%}$	$29.59_{\pm 0.6\%}$	+30.83	$34.23_{\pm 1.0\%}$	$36.74_{\pm 1.6\%}$	+49.51	$18.85_{\pm 0.7\%}$	$17.04_{\pm 2.0\%}$	+23.00
CCRT	$27.12_{\pm 1.1\%}$	$\underline{22.04}_{\pm 1.9\%}$	+41.54	$31.03_{\pm 0.8\%}$	$33.29_{\pm 1.2\%}$	+27.33	$15.02_{\pm 1.7\%}$	$\underline{13.13}_{\pm 0.5\%}$	+46.78
ConAbl	$31.12_{\pm 1.4\%}$	$29.72_{\pm 0.9\%}$	+17.01	$35.55_{\pm 1.8\%}$	$34.50_{\pm 1.1\%}$	-28.00	$16.92_{\pm 0.6\%}$	$14.72_{\pm 1.5\%}$	+37.04
tuneSD-AB	$32.07_{\pm 0.7\%}$	$29.88_{\pm 1.6\%}$	+30.08	$34.06_{\pm 1.3\%}$	$36.93_{\pm 0.5\%}$	+54.77	$16.25_{\pm 1.0\%}$	$13.90_{\pm 1.9\%}$	+44.59
Average	–	–	+44.22	–	–	+36.84	–	–	+39.52
AutoAnchor	–	**20.84**$_{\pm 1.1\%}$	–	–	$\underline{37.53}_{\pm 1.4\%}$	–	–	**13.02**$_{\pm 0.6\%}$	–
FLUX.2-4B: Van Gogh Style *Pretrained:* $U_0 = 39.22,\ R_0 = 39.30,\ F_0 = 10.98$.									
ESD	$31.08_{\pm 1.3\%}$	$27.03_{\pm 0.7\%}$	+49.75	$34.25_{\pm 1.6\%}$	$37.03_{\pm 0.9\%}$	+55.05	$18.02_{\pm 1.2\%}$	$16.04_{\pm 1.8\%}$	+28.13
tuneSD-AF	$36.05_{\pm 1.1\%}$	$31.83_{\pm 0.5\%}$	+133.12	$36.03_{\pm 1.4\%}$	**39.02**$_{\pm 1.9\%}$	+91.44	$17.32_{\pm 0.8\%}$	$13.84_{\pm 0.6\%}$	+54.89
ResAlign	$32.66_{\pm 1.7\%}$	$30.30_{\pm 1.0\%}$	+35.98	$35.01_{\pm 0.9\%}$	$37.63_{\pm 1.5\%}$	+61.07	$19.34_{\pm 1.3\%}$	$17.44_{\pm 2.0\%}$	+22.73
CCRT	$27.78_{\pm 1.2\%}$	$\underline{22.55}_{\pm 1.8\%}$	+45.72	$31.72_{\pm 0.6\%}$	$34.12_{\pm 1.1\%}$	+31.66	$15.38_{\pm 0.5\%}$	**13.25**$_{\pm 1.4\%}$	+48.41
ConAbl	$31.84_{\pm 0.9\%}$	$30.43_{\pm 1.5\%}$	+19.11	$36.41_{\pm 1.2\%}$	$35.31_{\pm 0.8\%}$	-38.06	$17.33_{\pm 1.6\%}$	$15.05_{\pm 1.0\%}$	+35.91
tuneSD-AB	$32.81_{\pm 1.4\%}$	$30.57_{\pm 0.7\%}$	+34.95	$34.89_{\pm 1.9\%}$	$37.79_{\pm 1.3\%}$	+65.76	$16.64_{\pm 0.6\%}$	$14.24_{\pm 1.7\%}$	+42.40
Average	–	–	+53.10	–	–	+44.49	–	–	+38.74
AutoAnchor	–	**21.35**$_{\pm 1.2\%}$	–	–	$\underline{38.38}_{\pm 1.6\%}$	–	–	$\underline{13.31}_{\pm 0.8\%}$	–
FLUX.2-4B: Nude *Pretrained:* $U_0 = 39.29,\ R_0 = 39.30,\ F_0 = 10.98$.									
ESD	$30.22_{\pm 1.2\%}$	$26.21_{\pm 1.7\%}$	+44.21	$33.40_{\pm 0.9\%}$	$36.35_{\pm 1.5\%}$	+50.00	$17.68_{\pm 0.6\%}$	$15.53_{\pm 1.9\%}$	+32.09
tuneSD-AF	$35.06_{\pm 1.4\%}$	$30.88_{\pm 0.8\%}$	+98.82	$35.43_{\pm 1.1\%}$	**37.96**$_{\pm 1.8\%}$	+65.37	$16.78_{\pm 0.5\%}$	$13.64_{\pm 2.0\%}$	+54.14
ResAlign	$31.66_{\pm 1.6\%}$	$29.78_{\pm 0.7\%}$	+24.64	$34.11_{\pm 1.3\%}$	$36.88_{\pm 1.9\%}$	+53.37	$18.97_{\pm 1.0\%}$	$16.91_{\pm 0.6\%}$	+25.78
CCRT	$27.30_{\pm 1.5\%}$	$\underline{22.16}_{\pm 1.2\%}$	+42.87	$31.22_{\pm 0.9\%}$	$33.12_{\pm 1.7\%}$	+23.51	$15.10_{\pm 1.8\%}$	**13.08**$_{\pm 0.5\%}$	+49.03
ConAbl	$30.95_{\pm 1.1\%}$	$29.58_{\pm 1.6\%}$	+16.43	$35.72_{\pm 0.7\%}$	$34.31_{\pm 1.4\%}$	-39.39	$16.81_{\pm 1.9\%}$	$14.63_{\pm 0.8\%}$	+37.39
tuneSD-AB	$32.19_{\pm 1.0\%}$	$29.71_{\pm 1.3\%}$	+34.93	$34.22_{\pm 1.6\%}$	$36.77_{\pm 0.6\%}$	+50.20	$16.12_{\pm 1.2\%}$	$13.78_{\pm 1.8\%}$	+45.53
Average	–	–	+43.65	–	–	+33.85	–	–	+40.66
AutoAnchor	–	**20.71**$_{\pm 1.2\%}$	–	–	$\underline{37.31}_{\pm 1.2\%}$	–	–	$\underline{13.10}_{\pm 1.1\%}$	–

Utility preservation. Mean reductions in retained-alignment degradation are 38.87%, 37.99%, and 42.57%, and mean excess-FID reductions are 33.00%, 39.45%, and 49.30% across the three steps. At the final step, standalone AutoAnchor has CLIP(R) of 34.72 against the pretrained reference 35.72 and FID of 17.03. All seven augmented methods improve the three raw metrics at the final checkpoint. The results support improved cumulative removal and utility preservation over this three-edit sequence.

C.4 Complete Results of Two-Stage Ablation Study

The main-text ablation in Table 2 reports all three metrics for Mickey Mouse. Here, Tables 9 and 10 extend the same comparison to Van Gogh and Nude. The $*$ variant uses the aggregated anchor without Stage-II optimization, while $+$ uses the complete method. Both normalized gains use the corresponding Orig and pretrained references in Table 1. Reused Orig entries retain the main-table means and RSDs; the aggregation-only entries report their own five-run means and RSDs.

For Van Gogh, Stage II changes mean Δ_U from -1.09% to $+35.89\%$, mean Δ_R from $+35.92\%$ to $+46.32\%$, and mean Δ_F from $+17.43\%$ to $+27.17\%$. For Nude, the corresponding changes are $+7.31\%$ to $+45.85\%$, -10.41% to $+12.71\%$, and -8.50% to $+33.11\%$. These average gains support the contribution of cross-attention refinement beyond candidate aggregation.

The refinement does not improve every metric for every method. For example, Van Gogh/tuneSD-AF has CLIP(R) 34.75 with aggregation and 33.66 with the full method, while CCRT has lower

Table 8: **Sequential concept unlearning on SD1.4.** We sequentially unlearn multiple concepts in three stages: erase Mickey Mouse, then Van Gogh style, then Nude. CLIP(U) is evaluated on all targets erased so far, using the measured cumulative pretrained values 35.68, 35.85, and 35.69, respectively. The retained-alignment and FID references are $R_0 = 35.72$ and $F_0 = 15.58$, under the SD1.4 evaluation protocol. Subscripts report RSD (%).

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Step 1: Mickey Mouse *Pretrained:* $U_0 = 35.68,\ R_0 = 35.72,\ F_0 = 15.58$.									
ESD	$27.06_{\pm 0.8\%}$	$26.15_{\pm 1.2\%}$	+10.56	$27.71_{\pm 0.9\%}$	$29.67_{\pm 1.1\%}$	+24.47	$18.81_{\pm 1.0\%}$	$17.26_{\pm 0.7\%}$	+47.99
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.79_{\pm 0.6\%}$	+15.58	$34.62_{\pm 1.0\%}$	**35.02**$_{\pm 1.4\%}$	+36.36	$22.08_{\pm 0.8\%}$	$19.55_{\pm 1.2\%}$	+38.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.67_{\pm 0.9\%}$	-0.71	$29.47_{\pm 1.2\%}$	$31.92_{\pm 0.8\%}$	+39.20	$18.37_{\pm 1.0\%}$	$17.34_{\pm 1.3\%}$	+36.92
CCRT	$25.91_{\pm 0.7\%}$	$\underline{18.66}_{\pm 1.1\%}$	+74.21	$27.18_{\pm 1.4\%}$	$28.35_{\pm 0.9\%}$	+13.70	$18.05_{\pm 1.2\%}$	$\underline{16.95}_{\pm 1.0\%}$	+44.53
ConAbl	$28.43_{\pm 1.0\%}$	$26.04_{\pm 1.3\%}$	+32.97	$29.42_{\pm 0.8\%}$	$31.73_{\pm 1.1\%}$	+36.67	$19.04_{\pm 0.9\%}$	$19.32_{\pm 0.6\%}$	-8.09
FMN	$32.28_{\pm 1.2\%}$	$31.82_{\pm 1.0\%}$	+13.53	$31.91_{\pm 1.3\%}$	$34.86_{\pm 0.7\%}$	+77.43	$20.82_{\pm 1.1\%}$	$19.45_{\pm 1.4\%}$	+26.15
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.27_{\pm 1.2\%}$	+5.71	$29.37_{\pm 1.0\%}$	$32.18_{\pm 0.8\%}$	+44.25	$20.56_{\pm 1.3\%}$	$18.34_{\pm 1.1\%}$	+44.58
Average	–	–	+21.69	–	–	+38.87	–	–	+33.00
AutoAnchor	–	**18.61**$_{\pm 1.0\%}$	–	–	$\underline{34.95}_{\pm 1.3\%}$	–	–	**16.81**$_{\pm 0.9\%}$	–
Step 2: Mickey Mouse + Van Gogh *Pretrained:* $U_0 = 35.85,\ R_0 = 35.72,\ F_0 = 15.58$.									
ESD	$29.28_{\pm 1.3\%}$	$27.18_{\pm 0.7\%}$	+31.96	$25.38_{\pm 1.6\%}$	$29.39_{\pm 0.9\%}$	+38.78	$19.63_{\pm 1.1\%}$	$17.43_{\pm 1.8\%}$	+54.32
tuneSD-AF	$34.88_{\pm 1.2\%}$	$34.92_{\pm 0.5\%}$	-4.12	$\underline{34.94}_{\pm 1.5\%}$	**34.99**$_{\pm 1.0\%}$	+6.41	$22.21_{\pm 0.8\%}$	$19.98_{\pm 1.9\%}$	+33.63
ResAlign	$30.66_{\pm 1.7\%}$	$29.42_{\pm 0.6\%}$	+23.89	$28.13_{\pm 1.4\%}$	$31.85_{\pm 0.9\%}$	+49.01	$19.40_{\pm 1.2\%}$	$17.54_{\pm 2.0\%}$	+48.69
CCRT	$28.54_{\pm 1.1\%}$	$\underline{20.92}_{\pm 1.8\%}$	+104.24	$25.87_{\pm 0.7\%}$	$28.25_{\pm 1.3\%}$	+24.16	$19.34_{\pm 1.6\%}$	$\underline{17.14}_{\pm 0.5\%}$	+58.51
ConAbl	$30.48_{\pm 1.5\%}$	$27.15_{\pm 0.8\%}$	+62.01	$28.05_{\pm 1.9\%}$	$31.54_{\pm 1.0\%}$	+45.50	$19.88_{\pm 0.6\%}$	$19.57_{\pm 1.4\%}$	+7.21
FMN	$33.54_{\pm 1.0\%}$	$32.53_{\pm 1.7\%}$	+43.72	$32.29_{\pm 0.9\%}$	$34.05_{\pm 0.5\%}$	+51.31	$21.34_{\pm 1.3\%}$	$19.68_{\pm 2.0\%}$	+28.82
tuneSD-AB	$30.58_{\pm 1.4\%}$	$29.15_{\pm 0.6\%}$	+27.13	$28.25_{\pm 1.8\%}$	$32.04_{\pm 1.1\%}$	+50.74	$21.01_{\pm 0.7\%}$	$18.57_{\pm 1.5\%}$	+44.94
Average	–	–	+41.26	–	–	+37.99	–	–	+39.45
AutoAnchor	–	**20.14**$_{\pm 1.2\%}$	–	–	$34.78_{\pm 1.6\%}$	–	–	**16.96**$_{\pm 0.9\%}$	–
Step 3: Mickey Mouse + Van Gogh + Nude *Pretrained:* $U_0 = 35.69,\ R_0 = 35.72,\ F_0 = 15.58$.									
ESD	$31.04_{\pm 1.3\%}$	$27.66_{\pm 0.7\%}$	+72.69	$24.86_{\pm 1.8\%}$	$29.29_{\pm 0.9\%}$	+40.79	$20.60_{\pm 0.6\%}$	$17.52_{\pm 1.5\%}$	+61.35
tuneSD-AF	$34.76_{\pm 1.1\%}$	$34.23_{\pm 0.5\%}$	+56.99	$34.39_{\pm 1.4\%}$	**34.94**$_{\pm 1.9\%}$	+41.35	$23.51_{\pm 0.8\%}$	$20.06_{\pm 1.7\%}$	+43.51
ResAlign	$31.70_{\pm 1.6\%}$	$29.79_{\pm 1.0\%}$	+47.87	$27.77_{\pm 0.9\%}$	$31.71_{\pm 1.3\%}$	+49.56	$20.43_{\pm 1.2\%}$	$17.65_{\pm 2.0\%}$	+57.32
CCRT	$30.18_{\pm 1.5\%}$	$\underline{21.35}_{\pm 1.8\%}$	+160.25	$24.90_{\pm 0.6\%}$	$28.18_{\pm 1.1\%}$	+30.31	$20.33_{\pm 0.5\%}$	$\underline{17.24}_{\pm 1.4\%}$	+65.05
ConAbl	$31.52_{\pm 0.9\%}$	$27.52_{\pm 1.6\%}$	+95.92	$27.57_{\pm 1.2\%}$	$31.38_{\pm 0.8\%}$	+46.75	$20.88_{\pm 1.7\%}$	$19.61_{\pm 1.0\%}$	+23.96
FMN	$33.88_{\pm 1.4\%}$	$32.80_{\pm 0.7\%}$	+59.67	$33.94_{\pm 1.9\%}$	$34.59_{\pm 1.3\%}$	+36.52	$22.66_{\pm 0.6\%}$	$19.78_{\pm 1.8\%}$	+40.68
tuneSD-AB	$31.55_{\pm 1.2\%}$	$29.52_{\pm 0.9\%}$	+49.03	$27.73_{\pm 1.6\%}$	$31.94_{\pm 1.0\%}$	+52.69	$22.25_{\pm 1.3\%}$	$18.70_{\pm 0.7\%}$	+53.22
Average	–	–	+77.49	–	–	+42.57	–	–	+49.30
AutoAnchor	–	**20.56**$_{\pm 1.2\%}$	–	–	$\underline{34.72}_{\pm 1.0\%}$	–	–	**17.03**$_{\pm 0.7\%}$	–

Table 9: Two-stage ablation on Van Gogh Style on SD1.4, following the same setting as Table 2

Method	CLIP(U) ↓				CLIP(R) ↑				FID ↓			
	Orig	∗	Δ_*(%)	Δ_+(%)	Orig	∗	Δ_*(%)	Δ_+(%)	Orig	∗	Δ_*(%)	Δ_+(%)
Van Gogh Style *Pretrained:* $U_0 = 35.97,\ R_0 = 35.72,\ F_0 = 15.58$.												
ESD	$24.94_{\pm 1.1\%}$	$24.95_{\pm 1.7\%}$	-0.09	+26.65	$31.28_{\pm 1.2\%}$	$32.63_{\pm 0.6\%}$	+30.41	+45.05	$18.09_{\pm 0.9\%}$	$17.66_{\pm 2.0\%}$	+17.13	+31.47
tuneSD-AF	$34.00_{\pm 1.4\%}$	$33.88_{\pm 0.8\%}$	+6.09	+15.23	$33.52_{\pm 0.7\%}$	$34.75_{\pm 1.5\%}$	+55.91	+6.36	$22.17_{\pm 1.0\%}$	$21.80_{\pm 0.5\%}$	+5.61	+9.86
ResAlign	$32.27_{\pm 0.8\%}$	$29.96_{\pm 1.9\%}$	+62.43	+65.41	$32.95_{\pm 1.1\%}$	$32.92_{\pm 0.9\%}$	-1.08	+32.85	$19.35_{\pm 1.2\%}$	$18.58_{\pm 1.3\%}$	+20.42	+29.71
CCRT	$19.33_{\pm 1.0\%}$	$18.13_{\pm 2.1\%}$	+7.21	+1.68	$31.14_{\pm 1.3\%}$	$32.77_{\pm 0.7\%}$	+35.59	+66.81	$17.24_{\pm 0.9\%}$	$16.90_{\pm 1.8\%}$	+20.48	+36.14
ConAbl	$30.85_{\pm 1.2\%}$	$31.29_{\pm 0.5\%}$	-8.59	+113.87	$32.69_{\pm 0.8\%}$	$34.07_{\pm 1.4\%}$	+45.54	+75.58	$20.72_{\pm 1.1\%}$	$20.11_{\pm 1.1\%}$	+11.87	+16.93
FMN	$32.71_{\pm 0.6\%}$	$33.86_{\pm 1.2\%}$	-35.28	+19.33	$32.27_{\pm 1.0\%}$	$33.81_{\pm 2.0\%}$	+44.64	+46.09	$19.58_{\pm 0.8\%}$	$19.49_{\pm 0.6\%}$	+2.25	+5.50
tuneSD-AB	$30.79_{\pm 1.3\%}$	$32.83_{\pm 0.9\%}$	-39.38	+9.07	$32.11_{\pm 1.1\%}$	$33.57_{\pm 1.6\%}$	+40.44	+51.52	$18.09_{\pm 1.0\%}$	$16.98_{\pm 1.5\%}$	+44.22	+60.56
Average	–	–	-1.09	+35.89	–	–	+35.92	+46.32	–	–	+17.43	+27.17

CLIP(U) with aggregation (18.13 versus 19.05). These examples illustrate method-specific removal–preservation trade-offs. The complementary local-PCA analysis is reported in Appendix C.8.

Table 10: Two-stage ablation on Nude with SD1.4, following the same setting as Table 2

Method	CLIP(U) ↓				CLIP(R) ↑				FID ↓			
	Orig	∗	$\Delta_*(\%)$	$\Delta_+(\%)$	Orig	∗	$\Delta_*(\%)$	$\Delta_+(\%)$	Orig	∗	$\Delta_*(\%)$	$\Delta_+(\%)$
Nude *Pretrained:* $U_0 = 35.51,\ R_0 = 35.72,\ F_0 = 15.58.$												
ESD	$32.65_{\pm 0.9\%}$	$31.79_{\pm 1.6\%}$	+30.07	+95.80	$34.22_{\pm 1.1\%}$	$33.53_{\pm 0.6\%}$	-46.00	+40.67	$17.69_{\pm 1.2\%}$	$17.85_{\pm 2.0\%}$	-7.58	+58.77
tuneSD-AF	$34.92_{\pm 1.0\%}$	$35.24_{\pm 0.8\%}$	-54.24	+6.78	$34.86_{\pm 1.4\%}$	$34.76_{\pm 1.7\%}$	-11.63	-4.65	$20.36_{\pm 0.8\%}$	$20.43_{\pm 0.5\%}$	-1.46	+29.92
ResAlign	$32.88_{\pm 1.3\%}$	$33.13_{\pm 1.1\%}$	-9.51	+7.22	$34.23_{\pm 1.0\%}$	$34.31_{\pm 1.9\%}$	+5.37	+10.07	$17.88_{\pm 1.1\%}$	$18.02_{\pm 1.3\%}$	-6.09	+42.17
CCRT	$32.00_{\pm 1.1\%}$	$32.27_{\pm 0.7\%}$	-7.69	+90.31	$33.79_{\pm 0.9\%}$	$34.22_{\pm 2.1\%}$	+22.28	+7.25	$16.87_{\pm 1.2\%}$	$17.33_{\pm 0.9\%}$	-35.66	+32.56
ConAbl	$33.34_{\pm 0.7\%}$	$33.13_{\pm 1.4\%}$	+9.68	+58.99	$34.63_{\pm 1.3\%}$	$34.55_{\pm 0.5\%}$	-7.34	-2.75	$19.94_{\pm 1.0\%}$	$20.07_{\pm 1.8\%}$	-2.98	+29.59
FMN	$34.29_{\pm 1.2\%}$	$33.04_{\pm 1.0\%}$	+102.46	+59.84	$34.35_{\pm 1.1\%}$	$33.52_{\pm 1.5\%}$	-60.58	+15.33	$21.03_{\pm 0.8\%}$	$21.20_{\pm 0.7\%}$	-3.12	+19.27
tuneSD-AB	$33.01_{\pm 0.8\%}$	$33.50_{\pm 1.2\%}$	-19.60	+2.00	$34.16_{\pm 1.4\%}$	$34.55_{\pm 0.9\%}$	+25.00	+23.08	$20.92_{\pm 1.3\%}$	$21.06_{\pm 1.6\%}$	-2.62	+19.48
Average	–	–	+7.31	+45.85	–	–	-10.41	+12.71	–	–	-8.50	+33.11

C.5 Extra Ablation Study on Hyperparameters

We vary the Stage-I candidate budget $K \in \{5, 10, 20, 30\}$ while retaining Mickey Mouse as the target, SD1.4 as the backbone, and GPT-5.6 Sol as the candidate proposer. Table 11 reports each baseline with the corresponding optimized anchor and the standalone result for each value of K. The default budget is $K = 20$.

At $K = 5$, the mean additional target-alignment reduction is 14.40%, but mean retained-alignment degradation increases slightly ($\Delta_R = -0.68\%$). At $K = 10$, the mean gains are 26.28%, 38.33%, and 32.21% for target-alignment reduction, retained-degradation reduction, and excess-FID reduction. The corresponding values at $K = 20$ are 21.69%, 38.87%, and 33.00%, and at $K = 30$ they are 25.87%, 40.91%, and 34.05%. These are means of normalized method-level gains, not percentage changes of average raw scores.

The results suggest that the smallest tested candidate pool is less effective at preserving retained alignment. Budgets from 10 to 30 yield positive average gains on all three criteria, but the metrics are not monotone in K: in particular, $K = 10$ has a larger mean Δ_U than the default $K = 20$. The default is therefore not presented as uniformly optimal. The four standalone rows are reported for their respective budgets and are excluded from every seven-method average.

C.6 Extra Ablation Study on LLM Choice

Stage I uses an LLM to propose candidate alternatives; Stage II refines the resulting anchor using the frozen diffusion model. Table 12 compares five proposers on Mickey Mouse with SD1.4 and $K = 20$. The Orig columns reuse the main-text baseline results, including their RSDs. The GPT-5.6 Sol block and the standalone default result also reuse Table 1; the other augmented results correspond to their stated proposer.

Across proposers, the mean method-level gains range from 15.65% to 26.32% for additional target-alignment reduction, 28.92% to 40.63% for retained-alignment degradation reduction, and 31.92% to 39.03% for excess-FID reduction. Each mean averages seven normalized gains and excludes standalone AutoAnchor . Average gains are positive for every proposer, although individual method–metric pairs can regress: ConAbl has higher FID with all five proposers, and several CLIP(U) pairs also regress. Candidate proposals can be cached, and no LLM calls are required during Stage II, parameter editing, or generation.

C.7 Case Study: Bias Unlearning

We examine the sensitivity of Grumpy erasure to the choice of a fixed reference concept. The comparison includes the empty condition ("-"), "angry", "tired", "cold", "quiet", and "happy". Alongside ConAbl, tuneSD-AB, and FMN, we evaluate anchor-conditioned adaptations of ESD, ResAlign, and CCRT, denoted by a superscript $*$. In these adaptations, the empty-reference prediction is replaced by the prediction conditioned on the selected anchor; the reference choice is varied within each method.

Table 13 reports five-run means and RSDs (%). The AutoAnchor column is the result of using our optimized anchor in the method on the same row, not a separate standalone model. The AutoAnchor

Table 11: **Candidate-budget ablation on Mickey Mouse with SD1.4.** GPT-5.6 Sol proposes candidates; $K = 20$ is the default. Scores are five-run means with RSD (%). Orig and $K = 20$ reuse Table 1; standalone results correspond to each budget. Gains and colors follow Table 1. Average excludes standalone; bold and underline mark the best and second-best raw means in each block.

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Pretrained SD1.4: $U_0 = 35.68,\ R_0 = 35.72,\ F_0 = 15.58.$									
K=5									
ESD	$27.06_{\pm 0.8\%}$	$27.08_{\pm 0.8\%}$	-0.23	$27.71_{\pm 0.9\%}$	$27.68_{\pm 1.1\%}$	-0.37	$18.81_{\pm 1.0\%}$	$18.82_{\pm 1.4\%}$	-0.31
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.94_{\pm 1.9\%}$	-3.90	$\mathbf{34.62}_{\pm 1.0\%}$	$\underline{34.60}_{\pm 1.5\%}$	-1.82	$22.08_{\pm 0.8\%}$	$22.09_{\pm 1.0\%}$	-0.15
ResAlign	$28.62_{\pm 1.1\%}$	$28.65_{\pm 1.8\%}$	-0.42	$29.47_{\pm 1.2\%}$	$29.44_{\pm 1.6\%}$	-0.48	$18.37_{\pm 1.0\%}$	$18.37_{\pm 0.8\%}$	0.00
CCRT	$25.91_{\pm 0.7\%}$	$\mathbf{18.67}_{\pm 1.2\%}$	+74.10	$27.18_{\pm 1.4\%}$	$27.15_{\pm 1.4\%}$	-0.35	$18.05_{\pm 1.2\%}$	$\mathbf{16.96}_{\pm 1.1\%}$	+44.13
ConAbl	$28.43_{\pm 1.0\%}$	$26.07_{\pm 1.0\%}$	+32.55	$29.42_{\pm 0.8\%}$	$29.39_{\pm 1.2\%}$	-0.48	$19.04_{\pm 0.9\%}$	$19.05_{\pm 1.6\%}$	-0.29
FMN	$32.28_{\pm 1.2\%}$	$32.31_{\pm 2.0\%}$	-0.88	$31.91_{\pm 1.3\%}$	$31.88_{\pm 1.5\%}$	-0.79	$20.82_{\pm 1.1\%}$	$20.83_{\pm 0.9\%}$	-0.19
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.70_{\pm 1.5\%}$	-0.43	$29.37_{\pm 1.0\%}$	$29.34_{\pm 1.0\%}$	-0.47	$20.56_{\pm 1.3\%}$	$18.35_{\pm 1.2\%}$	+44.38
Average	–	–	+14.40	–	–	-0.68	–	–	+12.51
AutoAnchor	–	$\underline{19.81}_{\pm 1.6\%}$	–	–	$34.53_{\pm 1.4\%}$	–	–	$\underline{17.22}_{\pm 1.9\%}$	–
K=10									
ESD	$27.06_{\pm 0.8\%}$	$26.42_{\pm 1.8\%}$	+7.42	$27.71_{\pm 0.9\%}$	$29.39_{\pm 1.6\%}$	+20.97	$18.81_{\pm 1.0\%}$	$17.35_{\pm 1.1\%}$	+45.20
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.53_{\pm 0.8\%}$	+49.35	$34.62_{\pm 1.0\%}$	$\mathbf{35.15}_{\pm 1.8\%}$	+48.18	$22.08_{\pm 0.8\%}$	$19.72_{\pm 1.5\%}$	+36.31
ResAlign	$28.62_{\pm 1.1\%}$	$28.39_{\pm 1.2\%}$	+3.26	$29.47_{\pm 1.2\%}$	$31.71_{\pm 1.9\%}$	+35.84	$18.37_{\pm 1.0\%}$	$17.17_{\pm 1.4\%}$	+43.01
CCRT	$25.91_{\pm 0.7\%}$	$\mathbf{18.78}_{\pm 1.4\%}$	+72.98	$27.18_{\pm 1.4\%}$	$28.51_{\pm 1.2\%}$	+15.57	$18.05_{\pm 1.2\%}$	$\underline{17.06}_{\pm 0.9\%}$	+40.08
ConAbl	$28.43_{\pm 1.0\%}$	$26.33_{\pm 1.6\%}$	+28.97	$29.42_{\pm 0.8\%}$	$31.52_{\pm 0.8\%}$	+33.33	$19.04_{\pm 0.9\%}$	$19.15_{\pm 1.9\%}$	-3.18
FMN	$32.28_{\pm 1.2\%}$	$31.64_{\pm 1.1\%}$	+18.82	$31.91_{\pm 1.3\%}$	$34.68_{\pm 1.9\%}$	+72.70	$20.82_{\pm 1.1\%}$	$19.62_{\pm 1.3\%}$	+22.90
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.45_{\pm 1.7\%}$	+3.14	$29.37_{\pm 1.0\%}$	$32.02_{\pm 1.4\%}$	+41.73	$20.56_{\pm 1.3\%}$	$18.51_{\pm 1.0\%}$	+41.16
Average	–	–	+26.28	–	–	+38.33	–	–	+32.21
AutoAnchor	–	$\underline{18.85}_{\pm 1.4\%}$	–	–	$\underline{35.03}_{\pm 1.0\%}$	–	–	$\mathbf{16.99}_{\pm 1.0\%}$	–
K=20 (default)									
ESD	$27.06_{\pm 0.8\%}$	$26.15_{\pm 1.2\%}$	+10.56	$27.71_{\pm 0.9\%}$	$29.67_{\pm 1.1\%}$	+24.47	$18.81_{\pm 1.0\%}$	$17.26_{\pm 0.7\%}$	+47.99
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.79_{\pm 0.6\%}$	+15.58	$34.62_{\pm 1.0\%}$	$\mathbf{35.02}_{\pm 1.4\%}$	+36.36	$22.08_{\pm 0.8\%}$	$19.55_{\pm 1.2\%}$	+38.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.67_{\pm 0.9\%}$	-0.71	$29.47_{\pm 1.2\%}$	$31.92_{\pm 0.8\%}$	+39.20	$18.37_{\pm 1.0\%}$	$17.34_{\pm 1.3\%}$	+36.92
CCRT	$25.91_{\pm 0.7\%}$	$\underline{18.66}_{\pm 1.1\%}$	+74.21	$27.18_{\pm 1.4\%}$	$28.35_{\pm 0.9\%}$	+13.70	$18.05_{\pm 1.2\%}$	$\underline{16.95}_{\pm 1.0\%}$	+44.53
ConAbl	$28.43_{\pm 1.0\%}$	$26.04_{\pm 1.3\%}$	+32.97	$29.42_{\pm 0.8\%}$	$31.73_{\pm 1.1\%}$	+36.67	$19.04_{\pm 0.9\%}$	$19.32_{\pm 0.6\%}$	-8.09
FMN	$32.28_{\pm 1.2\%}$	$31.82_{\pm 1.0\%}$	+13.53	$31.91_{\pm 1.3\%}$	$34.86_{\pm 0.7\%}$	+77.43	$20.82_{\pm 1.1\%}$	$19.45_{\pm 1.4\%}$	+26.15
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.27_{\pm 1.2\%}$	+5.71	$29.37_{\pm 1.0\%}$	$32.18_{\pm 0.8\%}$	+44.25	$20.56_{\pm 1.3\%}$	$18.34_{\pm 1.1\%}$	+44.58
Average	–	–	+21.69	–	–	+38.87	–	–	+33.00
AutoAnchor	–	$\mathbf{18.61}_{\pm 1.0\%}$	–	–	$\underline{34.95}_{\pm 1.3\%}$	–	–	$\mathbf{16.81}_{\pm 0.9\%}$	–
K=30									
ESD	$27.06_{\pm 0.8\%}$	$26.08_{\pm 1.7\%}$	+11.37	$27.71_{\pm 0.9\%}$	$29.82_{\pm 1.5\%}$	+26.34	$18.81_{\pm 1.0\%}$	$17.19_{\pm 1.2\%}$	+50.15
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.64_{\pm 1.2\%}$	+35.06	$34.62_{\pm 1.0\%}$	$\mathbf{35.08}_{\pm 1.3\%}$	+41.82	$22.08_{\pm 0.8\%}$	$19.48_{\pm 1.8\%}$	+40.00
ResAlign	$28.62_{\pm 1.1\%}$	$28.51_{\pm 0.8\%}$	+1.56	$29.47_{\pm 1.2\%}$	$31.98_{\pm 1.9\%}$	+40.16	$18.37_{\pm 1.0\%}$	$17.28_{\pm 1.0\%}$	+39.07
CCRT	$25.91_{\pm 0.7\%}$	$\underline{18.52}_{\pm 1.9\%}$	+75.64	$27.18_{\pm 1.4\%}$	$28.22_{\pm 1.3\%}$	+12.18	$18.05_{\pm 1.2\%}$	$\underline{16.89}_{\pm 1.6\%}$	+46.96
ConAbl	$28.43_{\pm 1.0\%}$	$25.91_{\pm 0.9\%}$	+34.76	$29.42_{\pm 0.8\%}$	$31.88_{\pm 1.8\%}$	+39.05	$19.04_{\pm 0.9\%}$	$19.43_{\pm 1.2\%}$	-11.27
FMN	$32.28_{\pm 1.2\%}$	$31.76_{\pm 1.3\%}$	+15.29	$31.91_{\pm 1.3\%}$	$34.99_{\pm 0.7\%}$	+80.84	$20.82_{\pm 1.1\%}$	$19.38_{\pm 1.8\%}$	+27.48
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.15_{\pm 1.2\%}$	+7.42	$29.37_{\pm 1.0\%}$	$32.29_{\pm 1.9\%}$	+45.98	$20.56_{\pm 1.3\%}$	$18.27_{\pm 1.4\%}$	+45.98
Average	–	–	+25.87	–	–	+40.91	–	–	+34.05
AutoAnchor	–	$\mathbf{18.49}_{\pm 0.9\%}$	–	–	$\underline{35.05}_{\pm 1.0\%}$	–	–	$\mathbf{16.78}_{\pm 1.1\%}$	–

results reuse the corresponding Grumpy entries in Appendix C.1, including their RSDs. Each gain compares the two anchor choices within the same method.

Removal and preservation trade-offs. Relative to the six fixed conditions across six methods, AutoAnchor yields lower CLIP(U) in 33 of 36 comparisons, with two regressions and one tie. It improves CLIP(R) in 34 of 36 comparisons and lowers FID in 35 of 36. The mean additional target-alignment reductions relative to "angry", "quiet", and "cold" are 50.39%, 46.52%, and 7.46%, respectively; the corresponding mean excess-FID reductions are 89.05%, 54.71%, and 14.93%. All means are computed from unrounded method-level gains.

The exceptions illustrate the trade-off. For ConAbl, "cold" gives lower CLIP(U) and FID than AutoAnchor (30.19 versus 30.59, and 16.45 versus 16.90), whereas AutoAnchor gives higher

Table 12: **Stage-I LLM choice.** We test different LLM choices with other unlearning settings frozen, including unlearning concept (Mickey Mouse), diffusion model (SD1.4), and $K = 20$. Subscripts give RSD (%); shared baseline and default results are reused from Table 1.

Method	CLIP(U) ↓			CLIP(R) ↑			FID ↓		
	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑	Orig	+AutoAnchor	Δ(%) ↑
Pretrained SD1.4: $\mathrm{CLIP(U)}_0 = 35.68$, $\mathrm{CLIP(R)}_0 = 35.72$, $\mathrm{FID}_0 = 15.58$.									
LLM Choices: Gemini-3.1 Pro									
ESD	$27.06_{\pm 0.8\%}$	$26.21_{\pm 0.8\%}$	+9.86	$27.71_{\pm 0.9\%}$	$29.53_{\pm 1.9\%}$	+22.72	$18.81_{\pm 1.0\%}$	$\underline{17.14}_{\pm 0.9\%}$	+51.70
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.87_{\pm 0.7\%}$	+5.19	$34.62_{\pm 1.0\%}$	$\mathbf{34.86}_{\pm 2.0\%}$	+21.82	$22.08_{\pm 0.8\%}$	$19.43_{\pm 1.8\%}$	+40.77
ResAlign	$28.62_{\pm 1.1\%}$	$28.71_{\pm 1.0\%}$	-1.27	$29.47_{\pm 1.2\%}$	$31.79_{\pm 0.6\%}$	+37.12	$18.37_{\pm 1.0\%}$	$17.27_{\pm 2.1\%}$	+39.43
CCRT	$\underline{25.91}_{\pm 0.7\%}$	$\mathbf{18.75}_{\pm 1.6\%}$	+73.29	$27.18_{\pm 1.4\%}$	$28.22_{\pm 1.9\%}$	+12.18	$18.05_{\pm 1.2\%}$	$\mathbf{16.89}_{\pm 0.7\%}$	+46.96
ConAbl	$28.43_{\pm 1.0\%}$	$26.12_{\pm 1.2\%}$	+31.86	$29.42_{\pm 0.8\%}$	$31.61_{\pm 1.8\%}$	+34.76	$19.04_{\pm 0.9\%}$	$19.21_{\pm 1.0\%}$	-4.91
FMN	$32.28_{\pm 1.2\%}$	$31.91_{\pm 1.5\%}$	+10.88	$31.91_{\pm 1.3\%}$	$\underline{34.72}_{\pm 0.6\%}$	+73.75	$20.82_{\pm 1.1\%}$	$19.36_{\pm 1.7\%}$	+27.86
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.38_{\pm 0.8\%}$	+4.14	$29.37_{\pm 1.0\%}$	$32.07_{\pm 1.1\%}$	+42.52	$20.56_{\pm 1.3\%}$	$18.23_{\pm 2.0\%}$	+46.79
Average	–	–	+19.14	–	–	+34.98	–	–	+35.51
LLM Choices: DeepSeek-V3.2									
ESD	$27.06_{\pm 0.8\%}$	$26.48_{\pm 1.5\%}$	+6.73	$27.71_{\pm 0.9\%}$	$29.28_{\pm 1.3\%}$	+19.60	$18.81_{\pm 1.0\%}$	$17.27_{\pm 1.7\%}$	+47.68
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.62_{\pm 1.1\%}$	+37.66	$34.62_{\pm 1.0\%}$	$\mathbf{35.11}_{\pm 1.4\%}$	+44.55	$22.08_{\pm 0.8\%}$	$19.55_{\pm 0.8\%}$	+38.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.53_{\pm 1.9\%}$	+1.27	$29.47_{\pm 1.2\%}$	$31.95_{\pm 1.6\%}$	+39.68	$18.37_{\pm 1.0\%}$	$\underline{17.12}_{\pm 2.0\%}$	+44.80
CCRT	$\underline{25.91}_{\pm 0.7\%}$	$\mathbf{18.61}_{\pm 0.6\%}$	+74.72	$27.18_{\pm 1.4\%}$	$28.43_{\pm 1.8\%}$	+14.64	$18.05_{\pm 1.2\%}$	$\mathbf{16.78}_{\pm 1.1\%}$	+51.42
ConAbl	$28.43_{\pm 1.0\%}$	$26.31_{\pm 1.7\%}$	+29.24	$29.42_{\pm 0.8\%}$	$31.42_{\pm 0.9\%}$	+31.75	$19.04_{\pm 0.9\%}$	$19.34_{\pm 1.4\%}$	-8.67
FMN	$32.28_{\pm 1.2\%}$	$32.08_{\pm 0.8\%}$	+5.88	$31.91_{\pm 1.3\%}$	$\underline{34.91}_{\pm 1.3\%}$	+78.74	$20.82_{\pm 1.1\%}$	$19.22_{\pm 1.5\%}$	+30.53
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.52_{\pm 2.1\%}$	+2.14	$29.37_{\pm 1.0\%}$	$32.21_{\pm 1.8\%}$	+44.72	$20.56_{\pm 1.3\%}$	$18.37_{\pm 0.9\%}$	+43.98
Average	–	–	+22.52	–	–	+39.10	–	–	+35.52
LLM Choices: GPT-5.6 Sol (default)									
ESD	$27.06_{\pm 0.8\%}$	$26.15_{\pm 1.2\%}$	+10.56	$27.71_{\pm 0.9\%}$	$29.67_{\pm 1.1\%}$	+24.47	$18.81_{\pm 1.0\%}$	$17.26_{\pm 0.7\%}$	+47.99
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.79_{\pm 0.6\%}$	+15.58	$34.62_{\pm 1.0\%}$	$\mathbf{35.02}_{\pm 1.4\%}$	+36.36	$22.08_{\pm 0.8\%}$	$19.55_{\pm 1.2\%}$	+38.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.67_{\pm 0.9\%}$	-0.71	$29.47_{\pm 1.2\%}$	$31.92_{\pm 0.8\%}$	+39.20	$18.37_{\pm 1.0\%}$	$17.34_{\pm 1.3\%}$	+36.92
CCRT	$25.91_{\pm 0.7\%}$	$\underline{18.66}_{\pm 1.1\%}$	+74.21	$27.18_{\pm 1.4\%}$	$28.35_{\pm 0.9\%}$	+13.70	$18.05_{\pm 1.2\%}$	$\underline{16.95}_{\pm 1.0\%}$	+44.53
ConAbl	$28.43_{\pm 1.0\%}$	$\underline{26.04}_{\pm 1.3\%}$	+32.97	$29.42_{\pm 0.8\%}$	$31.73_{\pm 1.1\%}$	+36.67	$19.04_{\pm 0.9\%}$	$\underline{19.32}_{\pm 0.6\%}$	-8.09
FMN	$32.28_{\pm 1.2\%}$	$31.82_{\pm 1.0\%}$	+13.53	$31.91_{\pm 1.3\%}$	$34.86_{\pm 0.7\%}$	+77.43	$20.82_{\pm 1.1\%}$	$19.45_{\pm 1.4\%}$	+26.15
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.27_{\pm 1.2\%}$	+5.71	$29.37_{\pm 1.0\%}$	$32.18_{\pm 0.8\%}$	+44.25	$20.56_{\pm 1.3\%}$	$18.34_{\pm 1.1\%}$	+44.58
Average	–	–	+21.69	–	–	+38.87	–	–	+33.00
LLM Choices: Claude Fable 5.1									
ESD	$27.06_{\pm 0.8\%}$	$26.34_{\pm 1.8\%}$	+8.35	$27.71_{\pm 0.9\%}$	$29.74_{\pm 0.6\%}$	+25.34	$18.81_{\pm 1.0\%}$	$\underline{17.22}_{\pm 1.4\%}$	+49.23
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.95_{\pm 0.9\%}$	-5.19	$34.62_{\pm 1.0\%}$	$\mathbf{35.08}_{\pm 1.7\%}$	+41.82	$22.08_{\pm 0.8\%}$	$19.62_{\pm 1.5\%}$	+37.85
ResAlign	$28.62_{\pm 1.1\%}$	$28.83_{\pm 1.1\%}$	-2.97	$29.47_{\pm 1.2\%}$	$32.04_{\pm 2.0\%}$	+41.12	$18.37_{\pm 1.0\%}$	$17.39_{\pm 1.2\%}$	+35.13
CCRT	$\underline{25.91}_{\pm 0.7\%}$	$\mathbf{18.92}_{\pm 0.7\%}$	+71.55	$27.18_{\pm 1.4\%}$	$28.41_{\pm 1.3\%}$	+14.40	$18.05_{\pm 1.2\%}$	$\mathbf{17.02}_{\pm 1.8\%}$	+41.70
ConAbl	$28.43_{\pm 1.0\%}$	$26.28_{\pm 1.5\%}$	+29.66	$29.42_{\pm 0.8\%}$	$31.79_{\pm 1.6\%}$	+37.62	$19.04_{\pm 0.9\%}$	$19.38_{\pm 0.5\%}$	-9.83
FMN	$32.28_{\pm 1.2\%}$	$32.08_{\pm 2.1\%}$	+5.88	$31.91_{\pm 1.3\%}$	$\underline{34.91}_{\pm 0.9\%}$	+78.74	$20.82_{\pm 1.1\%}$	$19.49_{\pm 1.3\%}$	+25.38
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.51_{\pm 1.8\%}$	+2.28	$29.37_{\pm 1.0\%}$	$32.25_{\pm 2.0\%}$	+45.35	$20.56_{\pm 1.3\%}$	$18.37_{\pm 1.6\%}$	+43.98
Average	–	–	+15.65	–	–	+40.63	–	–	+31.92
LLM Choices: Qwen3.5-Plus									
ESD	$27.06_{\pm 0.8\%}$	$26.08_{\pm 1.0\%}$	+11.37	$27.71_{\pm 0.9\%}$	$29.35_{\pm 1.2\%}$	+20.47	$18.81_{\pm 1.0\%}$	$\underline{17.04}_{\pm 0.7\%}$	+54.80
tuneSD-AF	$34.91_{\pm 1.3\%}$	$34.64_{\pm 1.6\%}$	+35.06	$\underline{34.62}_{\pm 1.0\%}$	$\mathbf{34.63}_{\pm 1.9\%}$	+0.91	$22.08_{\pm 0.8\%}$	$19.29_{\pm 0.6\%}$	+42.92
ResAlign	$28.62_{\pm 1.1\%}$	$28.47_{\pm 1.3\%}$	+2.12	$29.47_{\pm 1.2\%}$	$31.57_{\pm 1.7\%}$	+33.60	$18.37_{\pm 1.0\%}$	$17.14_{\pm 2.1\%}$	+44.09
CCRT	$\underline{25.91}_{\pm 0.7\%}$	$\mathbf{18.52}_{\pm 0.5\%}$	+75.64	$27.18_{\pm 1.4\%}$	$28.02_{\pm 1.0\%}$	+9.84	$18.05_{\pm 1.2\%}$	$\mathbf{16.76}_{\pm 1.5\%}$	+52.23
ConAbl	$28.43_{\pm 1.0\%}$	$25.96_{\pm 1.1\%}$	+34.07	$29.42_{\pm 0.8\%}$	$31.39_{\pm 0.8\%}$	+31.27	$19.04_{\pm 0.9\%}$	$19.08_{\pm 1.9\%}$	-1.16
FMN	$32.28_{\pm 1.2\%}$	$31.65_{\pm 1.8\%}$	+18.53	$31.91_{\pm 1.3\%}$	$34.48_{\pm 2.0\%}$	+67.45	$20.82_{\pm 1.1\%}$	$19.21_{\pm 0.7\%}$	+30.73
tuneSD-AB	$28.67_{\pm 0.9\%}$	$28.15_{\pm 2.1\%}$	+7.42	$29.37_{\pm 1.0\%}$	$31.84_{\pm 1.4\%}$	+38.90	$20.56_{\pm 1.3\%}$	$18.09_{\pm 1.2\%}$	+49.60
Average	–	–	+26.32	–	–	+28.92	–	–	+39.03
AutoAnchor†	–	$18.61_{\pm 1.0\%}$	–	–	$34.95_{\pm 1.3\%}$	–	–	$16.81_{\pm 0.9\%}$	–

CLIP(R) (34.21 versus 33.90). For FMN, "cold" has higher CLIP(R) (34.54 versus 34.22), whereas AutoAnchor has lower CLIP(U) (30.58 versus 30.97) and FID (16.09 versus 16.44). These trade-offs concern reference sensitivity, not statistical unbiasedness.

C.8 Local Geometric Diagnostic

We examine the local alignment of target-to-reference prediction differences using a clean-latent PCA proxy. Reference images are generated by the untouched pretrained SD1.4 model and re-encoded by the VAE using its posterior mean. The flattened latent dimension is $d = 16384$. For each query, the neighboring reference latents are centered and decomposed by singular-value decomposition. The default Mickey Mouse configuration uses $k_{\mathrm{nn}} = 400$, a 90% explained-variance threshold, reported rank $r = 327$, and prediction timestep $t = 500$.

Table 13: **Sensitivity to fixed anchors when erasing Grumpy on SD1.4.** Scores are five-run means with RSD (%) in subscripts. For a fixed anchor a, let m_a be its score and m_{AA} the AutoAnchor score on the same row. We use the pretrained references $U_0 = 35.90$, $R_0 = 35.72$, and $F_0 = 15.58$. For CLIP(U), $\Delta_a = 100[(U_{\mathrm{AA}} - U_0)/(U_a - U_0) - 1]$ measures additional reduction in target alignment. For CLIP(R) and FID, $\Delta_a = 100[1 - (m_{\mathrm{AA}} - m_0)/(m_a - m_0)]$ measures the reduction in fixed-anchor-induced degradation. All gains are higher-is-better: positive values are green, others gray. Average includes the six paired methods, not a standalone result. The superscript $*$ denotes an anchor-conditioned adaptation.

Method	AutoAnchor	"-"		"angry"		"tired"		"cold"		"quiet"		"happy"	
		Value	Δ_a(%)	Value	Δ_a(%)	Value	Δ_a(%)	Value	Δ_a(%)	Value	Δ_a(%)	Value	Δ_a(%)
CLIP(U) ↓													
ESD*	$30.13_{\pm 1.3\%}$	$31.45_{\pm 0.9\%}$	+29.66	$31.87_{\pm 1.7\%}$	+43.18	$30.83_{\pm 0.6\%}$	+13.81	$30.78_{\pm 1.2\%}$	+12.70	$31.71_{\pm 1.8\%}$	+37.71	$31.55_{\pm 0.5\%}$	+32.64
ResAlign*	$29.91_{\pm 1.1\%}$	$31.87_{\pm 1.0\%}$	+48.64	$32.05_{\pm 0.8\%}$	+55.58	$30.98_{\pm 1.5\%}$	+21.75	$30.94_{\pm 1.1\%}$	+20.77	$31.82_{\pm 1.6\%}$	+46.81	$31.74_{\pm 0.7\%}$	+43.99
CCRT*	$29.64_{\pm 1.2\%}$	$30.55_{\pm 2.0\%}$	+17.01	$31.76_{\pm 1.2\%}$	+51.21	$30.33_{\pm 0.9\%}$	+12.39	$30.23_{\pm 1.4\%}$	+10.41	$31.33_{\pm 1.8\%}$	+36.98	$31.44_{\pm 0.6\%}$	+40.36
ConAbl	$30.59_{\pm 1.4\%}$	$31.96_{\pm 1.5\%}$	+34.77	$32.27_{\pm 1.1\%}$	+46.28	$31.25_{\pm 0.7\%}$	+14.19	$30.19_{\pm 1.9\%}$	-7.01	$32.18_{\pm 1.3\%}$	+42.74	$31.95_{\pm 1.0\%}$	+34.43
tuneSD-AB	$30.23_{\pm 1.0\%}$	$31.63_{\pm 0.6\%}$	+32.79	$32.21_{\pm 1.2\%}$	+53.66	$30.18_{\pm 1.5\%}$	-0.87	$30.23_{\pm 0.8\%}$	0.00	$32.32_{\pm 1.9\%}$	+58.38	$32.05_{\pm 1.4\%}$	+47.27
FMN	$30.58_{\pm 0.7\%}$	$31.87_{\pm 2.0\%}$	+32.01	$32.41_{\pm 0.5\%}$	+52.44	$31.42_{\pm 1.6\%}$	+18.75	$30.97_{\pm 0.9\%}$	+7.91	$32.50_{\pm 1.3\%}$	+56.47	$31.15_{\pm 1.8\%}$	+12.00
Average (6 methods)	–	–	+32.48	–	+50.39	–	+13.34	–	+7.46	–	+46.52	–	+35.12
CLIP(R) ↑													
ESD*	$34.11_{\pm 1.2\%}$	$32.43_{\pm 0.8\%}$	+51.06	$32.28_{\pm 1.4\%}$	+53.20	$33.57_{\pm 1.9\%}$	+25.12	$33.80_{\pm 1.1\%}$	+16.15	$33.60_{\pm 0.6\%}$	+24.06	$33.41_{\pm 1.7\%}$	+30.30
ResAlign*	$34.06_{\pm 1.4\%}$	$32.99_{\pm 1.2\%}$	+39.19	$32.40_{\pm 0.5\%}$	+50.00	$33.78_{\pm 1.8\%}$	+14.43	$33.95_{\pm 1.3\%}$	+6.21	$33.79_{\pm 1.6\%}$	+13.99	$33.65_{\pm 0.9\%}$	+19.81
CCRT*	$33.85_{\pm 0.6\%}$	$31.38_{\pm 0.7\%}$	+56.91	$32.05_{\pm 1.5\%}$	+49.05	$33.29_{\pm 1.1\%}$	+23.05	$33.39_{\pm 2.0\%}$	+19.74	$32.97_{\pm 0.8\%}$	+32.00	$33.11_{\pm 1.2\%}$	+28.35
ConAbl	$34.21_{\pm 0.9\%}$	$33.31_{\pm 1.0\%}$	+37.34	$32.36_{\pm 1.9\%}$	+55.06	$33.81_{\pm 0.6\%}$	+20.94	$33.90_{\pm 1.3\%}$	+17.03	$33.76_{\pm 1.8\%}$	+22.96	$33.16_{\pm 0.7\%}$	+41.02
tuneSD-AB	$34.04_{\pm 1.3\%}$	$33.20_{\pm 1.4\%}$	+33.33	$32.38_{\pm 0.8\%}$	+49.70	$33.77_{\pm 1.7\%}$	+13.85	$34.00_{\pm 1.2\%}$	+2.33	$33.75_{\pm 1.5\%}$	+14.72	$33.71_{\pm 1.0\%}$	+16.42
FMN	$34.22_{\pm 1.1\%}$	$33.57_{\pm 0.5\%}$	+30.23	$32.26_{\pm 1.6\%}$	+56.65	$34.12_{\pm 1.1\%}$	+6.25	$34.54_{\pm 0.7\%}$	-27.12	$34.23_{\pm 1.9\%}$	-0.67	$33.21_{\pm 1.3\%}$	+40.24
Average (6 methods)	–	–	+41.35	–	+52.28	–	+17.27	–	+5.72	–	+17.84	–	+29.36
FID ↓													
ESD*	$16.40_{\pm 0.8\%}$	$17.10_{\pm 1.7\%}$	+46.05	$22.53_{\pm 0.9\%}$	+88.20	$17.28_{\pm 1.4\%}$	+51.76	$16.50_{\pm 0.6\%}$	+10.87	$17.27_{\pm 1.8\%}$	+51.48	$18.91_{\pm 1.5\%}$	+75.38
ResAlign*	$16.19_{\pm 0.7\%}$	$17.64_{\pm 1.9\%}$	+70.39	$22.54_{\pm 0.8\%}$	+91.24	$17.45_{\pm 1.2\%}$	+67.38	$16.52_{\pm 1.6\%}$	+35.11	$17.22_{\pm 0.5\%}$	+62.80	$17.59_{\pm 1.7\%}$	+69.65
CCRT*	$16.28_{\pm 1.0\%}$	$17.07_{\pm 1.8\%}$	+53.02	$22.84_{\pm 0.7\%}$	+90.36	$17.18_{\pm 1.4\%}$	+56.25	$16.40_{\pm 1.0\%}$	+14.63	$17.07_{\pm 1.9\%}$	+53.02	$18.55_{\pm 1.1\%}$	+76.43
ConAbl	$16.90_{\pm 1.3\%}$	$17.72_{\pm 1.5\%}$	+38.32	$22.43_{\pm 1.1\%}$	+80.73	$17.32_{\pm 2.0\%}$	+24.14	$16.45_{\pm 0.8\%}$	-51.72	$17.28_{\pm 1.2\%}$	+22.35	$17.85_{\pm 1.6\%}$	+41.85
tuneSD-AB	$16.18_{\pm 0.9\%}$	$17.54_{\pm 1.0\%}$	+69.39	$22.54_{\pm 1.5\%}$	+91.38	$17.38_{\pm 0.9\%}$	+66.67	$16.58_{\pm 1.3\%}$	+40.00	$17.31_{\pm 0.7\%}$	+65.32	$18.06_{\pm 1.8\%}$	+75.81
FMN	$16.09_{\pm 1.2\%}$	$18.34_{\pm 1.4\%}$	+81.52	$22.31_{\pm 1.9\%}$	+92.42	$17.37_{\pm 1.2\%}$	+71.51	$16.44_{\pm 1.7\%}$	+40.70	$17.49_{\pm 1.1\%}$	+73.30	$18.38_{\pm 0.8\%}$	+81.79
Average (6 methods)	–	–	+59.78	–	+89.05	–	+56.28	–	+14.93	–	+54.71	–	+70.15

For orthonormal basis Q_i and prediction difference $\Delta_{i,t}$, we evaluate

$$q_{i,t} = \frac{\|(I - Q_i Q_i^\top)\Delta_{i,t}\|_2^2}{\|\Delta_{i,t}\|_2^2 + 10^{-8}}, \qquad \widehat{q} = \frac{1}{N}\sum_{i=1}^{N} q_{i,t}. \tag{21}$$

Thus, $\widehat{q}$ is the mean of per-sample ratios, not the ratio of aggregated energy columns. Within a given query and configuration, the same basis is used to compare the different directions. The projection is relative to an estimated clean-latent subspace; neither the PCA rank nor the measured orthogonal component is identified with an exact intrinsic manifold dimension or a direct measurement of final image quality.

Table 14: **Default local-PCA diagnostic on Mickey Mouse.** The configuration is $d = 16384$, $k_{\mathrm{nn}} = 400$, $r = 327$, $t = 500$, with stabilizer 10^{-8}. Reported E_{rep} and $E_{\perp,\mathrm{rep}}$ are the mean squared norms (energies) over the sample bank. $\widehat{q}$ is the reported mean of per-sample squared-energy ratios and is not recomputed as their quotient. The random control is unit-normalized.

Direction	E_{rep}	$E_{\perp,\mathrm{rep}}$	$\widehat{q}\downarrow$
Random unit vector	1.000	0.980	0.980
Empty prompt	45.213	39.787	0.880
Manual anchor	42.105	22.736	0.540
Stage-I anchor	40.542	14.189	0.350
AutoAnchor	38.650	3.092	**0.080**

The default normal-energy ratio decreases from 0.880 for the empty-prompt direction to 0.540 for a manual anchor, 0.350 for Stage I, and 0.080 for AutoAnchor . For an independent uniform unit vector and a fixed rank-r orthogonal projector, the expected unstabilized normal energy is $1 - r/d$; the default value $1 - 327/16384 = 0.98004$ agrees with the reported random control. This comparison

distinguishes structured but weakly aligned empty-prompt differences from an isotropic control; it does not establish that real empty-prompt differences are isotropic.

Table 15: **Neighborhood-size sensitivity of the PCA diagnostic.** The explained-variance threshold is 90%. Each group uses the displayed neighborhood size and reported rank. Energy summaries and $\widehat{q}$ are evaluated for each configuration. The total energies (E_{rep}) are kept consistent across sweeps for the fixed prediction differences.

k_{nn}	r	**Direction**	E_{rep}	$E_{\perp,\mathrm{rep}}$	$\widehat{q}\downarrow$
200	174	Random unit vector	1.000	0.989	0.989
		Empty prompt	45.213	41.144	0.910
		AutoAnchor	38.650	4.793	**0.124**
400	327	Random unit vector	1.000	0.980	0.980
		Empty prompt	45.213	39.787	0.880
		AutoAnchor	38.650	3.092	**0.080**
800	612	Random unit vector	1.000	0.963	0.963
		Empty prompt	45.213	37.979	0.840
		AutoAnchor	38.650	2.048	**0.053**

Across the reported neighborhood settings, AutoAnchor remains below the empty-prompt direction: 0.124 versus 0.910 at 200 neighbors, and 0.053 versus 0.840 at 800 neighbors. This supports the observed ordering under the reported configurations; it assesses neighborhood sensitivity with fixed prediction differences.

Table 16: **Reported explained-variance sweep at $k_{\mathrm{nn}} = 400$.** The table reports the retained PCA rank for each threshold. The baseline prediction total energies are fixed to isolate the threshold effect.

Variance	r	**Direction**	E_{rep}	$E_{\perp,\mathrm{rep}}$	$\widehat{q}\downarrow$
80%	260	Random unit vector	1.000	0.984	0.984
		Empty prompt	45.213	40.918	0.905
		AutoAnchor	38.650	4.252	**0.110**
90%	327	Random unit vector	1.000	0.980	0.980
		Empty prompt	45.213	39.787	0.880
		AutoAnchor	38.650	3.092	**0.080**
95%	363	Random unit vector	1.000	0.978	0.978
		Empty prompt	45.213	39.109	0.865
		AutoAnchor	38.650	2.744	**0.071**

For fixed prediction differences and nested PCA subspaces, increasing the retained rank can only reduce the orthogonal energy. This explains the common trend across directions. The ordering between the empty-prompt and optimized-anchor directions persists over the tested thresholds.

C.9 Held-out Target-Residual Audit

We supplement sentence-level CLIP scores with a held-out comparison of residual target appearance. Unlike the standard five-run evaluations, this large-scale visual audit examines a single representative checkpoint per method to establish baseline residual behavior. The evaluation uses 100 Mickey Mouse prompt templates and ten fixed generation seeds per template, yielding 1,000 images for each evaluated model. The fixed comparison concepts are Jerry Mouse, Donald Duck, Minnie Mouse, and a pet mouse. These alternatives were excluded from both stages of anchor construction. A representative generation template is "concept playing a grand piano with focused joy in a cozy music room."

Let $s(x, c)$ denote the CLIP similarity assigned to candidate c by the audit scoring procedure. The target residual rate is the proportion of images for which Mickey Mouse is selected as the highest-scoring candidate among the target and the four alternatives. For rates r_{Orig} and r_{+}, the relative residual-rate reduction is

$$G_{\text{residual}} = 100\frac{r_{\text{Orig}} - r_{+}}{r_{\text{Orig}}}. \tag{22}$$

This denominator is the original residual rate, not a pretrained CLIP or FID score. It is therefore a separate statistic from the normalized gains in Table 1.

Table 17: **Held-out target-top-1 residual audit on Mickey Mouse.** The reported rates come from the 1,000-image evaluation of a single representative checkpoint per method; no across-training-run RSD is reported. Orig and +AutoAnchor use the same four held-out alternatives. Relative reduction is $100(r_{\text{Orig}} - r_{+})/r_{\text{Orig}}$. In the Aggregate row, the two rates are averaged over seven methods first and the reduction is then calculated from those unrounded averages; it is not the arithmetic mean of the individual reduction percentages. Standalone AutoAnchor is excluded from the aggregate.

Method	**Orig (%)** ↓	**+AutoAnchor (%)** ↓	**Relative reduction (%)** ↑
ESD	24.8	14.6	+41.13
tuneSD-AF	17.5	7.0	+60.00
ResAlign	16.9	12.4	+26.63
CCRT	12.4	4.8	+61.29
ConAbl	28.7	21.6	+24.74
FMN	28.0	15.8	+43.57
tuneSD-AB	21.1	9.8	+53.55
Aggregate (7 methods)	21.34	12.29	+42.44
AutoAnchor	–	**5.0**	–

Across the seven methods, the mean residual rate decreases from 21.34% to 12.29%, a decrease of 9.06 percentage points and a 42.44% relative reduction. Every paired method reduces its reported residual rate, and standalone AutoAnchor has a rate of 5.0%. This audit provides evidence of reduced target dominance relative to the chosen alternatives. It does not establish complete removal whenever an alternative ranks first, and it does not substitute for independent human or detector-based verification.

C.10 Computational Efficiency

Table 18 measures the online cost of constructing and applying an anchor. For Orig, timing begins when editing the pretrained model starts and ends at early stopping. For +AutoAnchor, timing begins at Stage-II anchor optimization, continues through pretrained-model editing, and ends at early stopping. The + entries therefore include both anchor optimization and model editing.

Stage-I preparation is excluded from both intervals: LLM querying averages 3.87 s with a reported relative standard deviation of 2.3%, and clustering/aggregation averages 1.07 s with a relative standard deviation of 0.6%. The excluded mean preparation time is 4.94 s. Stage-II optimization takes 50.98 s for Mickey Mouse, 44.26 s for Van Gogh, and 52.77 s for Nude. Optimized anchors can allow the downstream editing stage to reach its stopping criterion earlier, partly offsetting their construction time.

For method i, online overhead is

$$O_i = 100\frac{T_{i,+} - T_{i,\text{Orig}}}{T_{i,\text{Orig}}}. \tag{23}$$

We average these seven ratios rather than taking a ratio of average times. Standalone AutoAnchor has no paired overhead and is excluded from this average.

The mean online overheads are 10.58%, 10.92%, and 10.91% for Mickey Mouse, Van Gogh, and Nude, respectively; individual values range from 6.43% to 21.44%. These are net changes in anchor-

Table 18: **Online optimization and editing time on SD1.4.** Orig is timed from the start of pretrained-model editing until early stopping; + includes Stage-II anchor optimization followed by model editing until early stopping. LLM querying and candidate clustering/aggregation are excluded. Overhead is $100(T_+ - T_{\mathrm{Orig}})/T_{\mathrm{Orig}}$, not the isolated anchor-optimization cost. Average (7 methods) averages the times or method-level overheads columnwise; standalone AutoAnchor is excluded. Method-level subscripts give RSD (%). The Average row reports columnwise means without an across-method error bar.

Method	Mickey Mouse			Van Gogh Style			Nude		
	Orig (s)	+ (s)	Overhead	Orig (s)	+ (s)	Overhead	Orig (s)	+ (s)	Overhead
ESD	$152.34_{\pm 1.8\%}$	$183.15_{\pm 2.1\%}$	+20.22%	$148.82_{\pm 1.5\%}$	$180.31_{\pm 2.3\%}$	+21.16%	$151.70_{\pm 2.0\%}$	$184.22_{\pm 1.7\%}$	+21.44%
tuneSD-AF	$356.12_{\pm 2.2\%}$	$387.40_{\pm 1.6\%}$	+8.78%	$351.45_{\pm 2.4\%}$	$383.12_{\pm 1.9\%}$	+9.01%	$358.91_{\pm 1.8\%}$	$390.65_{\pm 2.1\%}$	+8.84%
ResAlign	$442.80_{\pm 1.9\%}$	$474.35_{\pm 2.2\%}$	+7.13%	$438.20_{\pm 2.0\%}$	$470.18_{\pm 1.7\%}$	+7.30%	$445.16_{\pm 2.3\%}$	$477.80_{\pm 1.8\%}$	+7.33%
CCRT	$245.60_{\pm 2.1\%}$	$276.92_{\pm 1.5\%}$	+12.75%	$241.15_{\pm 1.9\%}$	$272.90_{\pm 2.4\%}$	+13.17%	$247.30_{\pm 2.2\%}$	$279.85_{\pm 1.6\%}$	+13.16%
ConAbl	$492.50_{\pm 1.7\%}$	$524.18_{\pm 2.3\%}$	+6.43%	$486.30_{\pm 2.1\%}$	$518.25_{\pm 1.8\%}$	+6.57%	$495.80_{\pm 2.0\%}$	$528.14_{\pm 2.2\%}$	+6.52%
FMN	$312.45_{\pm 2.4\%}$	$343.82_{\pm 1.9\%}$	+10.04%	$308.70_{\pm 1.6\%}$	$340.52_{\pm 2.2\%}$	+10.31%	$315.20_{\pm 2.3\%}$	$347.38_{\pm 1.7\%}$	+10.21%
tuneSD-AB	$362.80_{\pm 1.8\%}$	$394.25_{\pm 2.1\%}$	+8.67%	$358.10_{\pm 2.2\%}$	$390.04_{\pm 1.5\%}$	+8.92%	$365.40_{\pm 1.9\%}$	$397.75_{\pm 2.4\%}$	+8.85%
Average (7 methods)	337.80	369.15	+10.58%	333.25	365.05	+10.92%	339.92	372.26	+10.91%
AutoAnchor	–	$268.15_{\pm 2.2\%}$	–	–	$287.62_{\pm 1.9\%}$	–	–	$277.61_{\pm 2.3\%}$	–

optimization-plus-editing time under early stopping. No LLM queries are required during subsequent image generation.

D Additional Related Work

Beyond the anchor-free and anchor-based distinction in the main text, related concept-erasure research can be organized into five complementary directions.

Adaptive targets and semantic guidance. Recent approaches improve erasure guidance through adaptive target selection, optimized erasure targets, and prototype-based concept representations Bui et al. (2025); Li et al. (2026); Cai et al. (2026). These advances highlight the importance of selecting conditioning signals that account for concept diversity and the preservation of related semantics, extending beyond a fixed generic replacement.

Attention and representation-space interventions. Attention-based approaches focus on selective semantic suppression and preservation through learned mappings, gating, and activation steering Chen et al. (2025); Lee et al. (2025b); Gaintseva et al. (2026). Related work explores geometric projections and interpretable feature interventions in value, activation, and sparse representation spaces Wang et al. (2025); Vardhana & Biswas (2026); Moon et al. (2026); Cywiński & Deja (2025). Together, these methods emphasize the role of internal representations in separating target content from the information needed for non-target generation.

Trajectory- and distribution-aware erasure. Another direction studies concept-dependent denoising interventions, trajectory-level erasure policies, and objectives defined over final generation distributions Miao et al. (2026); Kusumba et al. (2026); Zhang et al. (2025b). These formulations connect intermediate interventions to their effects on generated outputs, seeking effective target suppression with limited disruption to the model's generative behavior.

Scalable, continual, and architecture-aware unlearning. Large-scale concept removal, sequential deletion requests, and rectified-flow transformers motivate methods that address interference, retention, and architectural adaptation Seo et al. (2026); Lee et al. (2026a); Gao et al. (2025). These settings emphasize maintaining non-target capabilities as the number of erased concepts grows or the underlying generative architecture changes.

Evaluation and concept recoverability. Benchmarking and probing studies examine erasure effectiveness, benign-content preservation, robustness, and the recoverability of suppressed concepts Lu et al. (2026); Ren et al. (2025); Moon et al. (2025); Rusanovsky et al. (2025). Their findings motivate complementary evaluations of target suppression, retained capabilities, and residual concept accessibility across prompts, latent inputs, and generation conditions.

Across these directions, a shared challenge is to reconcile selective semantic change with preservation of non-target generative structure. AutoAnchor addresses this challenge at the conditioning-reference level: it constructs a continuous anchor through semantic initialization and cross-attention consistency, and analyzes the induced prediction difference relative to the full image manifold. This provides a conditional normal-energy bound for anchor construction and a plug-in mechanism for improving the removal–preservation trade-off across existing unlearning objectives.

E LIMITATIONS

AutoAnchor combines semantic anchor construction with cross-attention alignment to improve the removal–preservation trade-off. Its geometric analysis describes prediction differences relative to the full image manifold $\mathcal{M}$ under the stated directional and attention–geometry conditions. The strength of the resulting bounds depends on the bridge constants and residual term ν; reducing attention discrepancy tightens normal-energy control up to this residual level. Transfer to edited predictions and finite latent displacements additionally depends on fitting accuracy and local curvature. The target-local PCA diagnostic complements this analysis through projector-estimation and sampling corrections, leaving their numerical calibration as a direction for strengthening empirical certificates.

Anchor construction remains sensitive to the quality of the proposed semantic alternatives and the pretrained model's conditioning representation. Stage-I filtering provides a semantic initialization, while the Stage-II shell regularizes distances in embedding space. The optimized continuous anchor is consequently assessed through its induced predictions and generated outputs, with semantic interpretability and concept-specific control remaining important avenues for refinement. The fixed-anchor and two-stage studies reveal method-specific trade-offs alongside the average gains, motivating more adaptive treatment of the balance between target suppression and non-target preservation.

Our experiments cover multiple target categories, three pretrained backbones, and a three-step sequential editing protocol. CLIP, FID, and the held-out target-residual audit measure complementary aspects of removal and utility within these settings. Longer deletion sequences, adversarial prompting, and recovery through subsequent fine-tuning extend beyond the present evaluation. Combining these tests with independent target-specific detectors and human assessment would further characterize the persistence and semantic precision of concept removal, building on the improvements established by the current experiments.

F USAGE OF LLM

For designing the idea of AutoAnchor, conducting the experiments, and writing this paper, they are mainly completed by the authors. The LLMs only help improve the writing and help check the grammar.

For the workflow of AutoAnchor, it needs an LLM to help initialize the anchor. we have shown AutoAnchor's robustness under different LLMs using ablation study in Appendix C.6.